INTELIGENCIA DE NEGOCIO
2019 - 2020

- **Tema 1. Introducción a la Inteligencia de Negocio**
- **Tema 2. Minería de Datos. Ciencia de Datos**
- **Tema 3. Modelos de Predicción: Clasificación, regresión y series temporales**
- **Tema 4. Preparación de Datos**
- **Tema 5. Modelos de Agrupamiento o Segmentación**
- **Tema 6. Modelos de Asociación**
- **Tema 7. Modelos Avanzados de Minería de Datos**
- **Tema 8. Big Data**

Modelos avanzados de Minería de Datos

Objetivos:

- Analizar diferentes problemas y técnicas de ciencia de datos, tanto extensiones del problema clasificación clásico con nuevos problemas: anomalías, flujo continuo de datos, análisis de sentimientos … técnicas como deep learning, ….

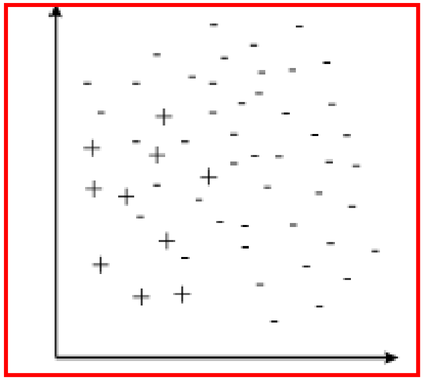
Inteligencia de Negocio

TEMA 7. Modelos Avanzados de Minería de Datos

1. **Clases no balanceadas/equilibradas**
2. Características intrínsecas de los datos en clasificación
3. Detección de anomalías
4. Problemas no estándar de clasificación: MIL, MLL, ...
5. Análisis de Sentimientos
6. Deep Learning

Classification with Imbalanced Data Sets Presentation

In a concept-learning problem, the data set is said to present a class imbalance if it contains many more examples of one class than the other.



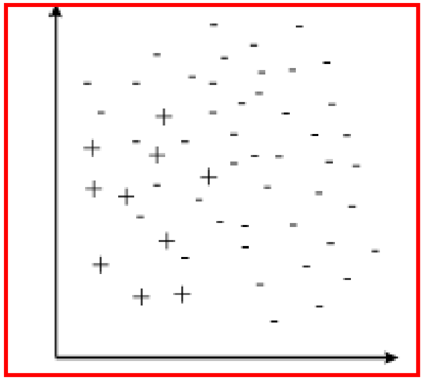
There exist many domains that do not have a balanced data set. There are a lot of problems where the most important knowledge usually resides in the minority class.

Ej.: Detection of uncommon diseases presents Imbalanced data: Few sick persons and lots of healthy persons.

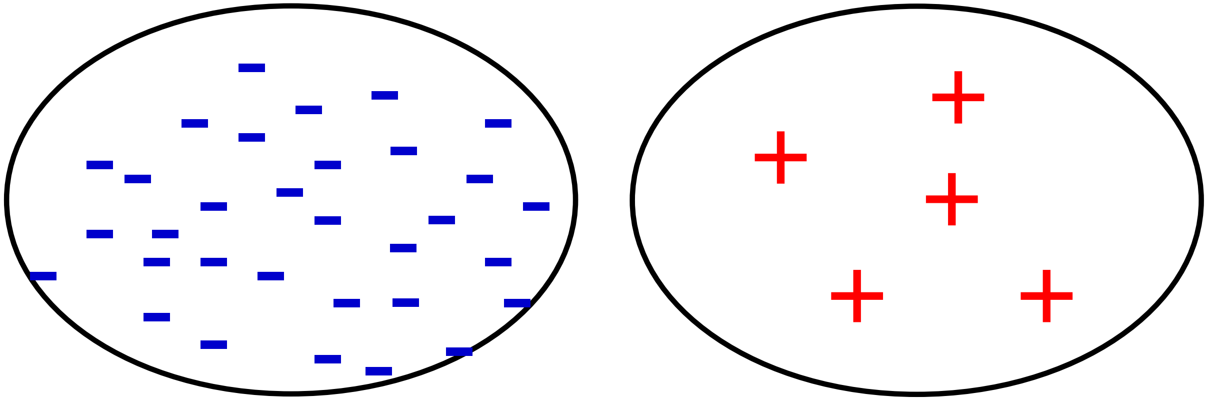
Some real-problems: Fraudulent credit card transactions, Learning word pronunciation, Prediction of telecommunications equipment failures, Detection oil spills from satellite images, Detection of Melanomas, Intrusion detection, Insurance risk modeling, Hardware fault detection

Classification with Imbalanced Data Sets Presentation

Such a situation introduce challenges for typical classifiers (such as decision tree) "systems that are designed to optimize overall accuracy without taking into account the relative distribution of each class".

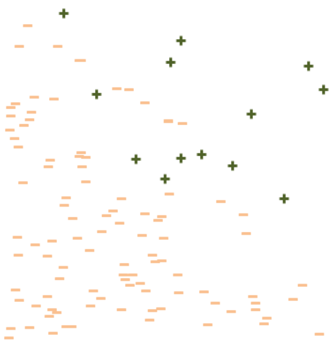


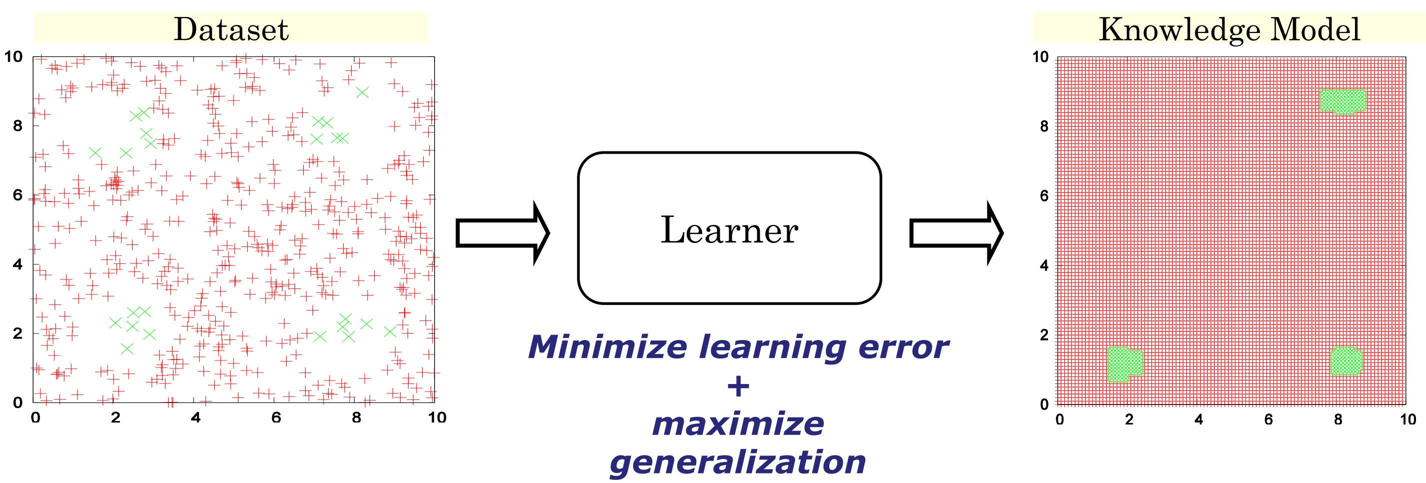
As a result, these classifiers tend to ignore small classes while concentrating on classifying the large ones accurately.



Why learning from imbalanced data-sets might be difficult?

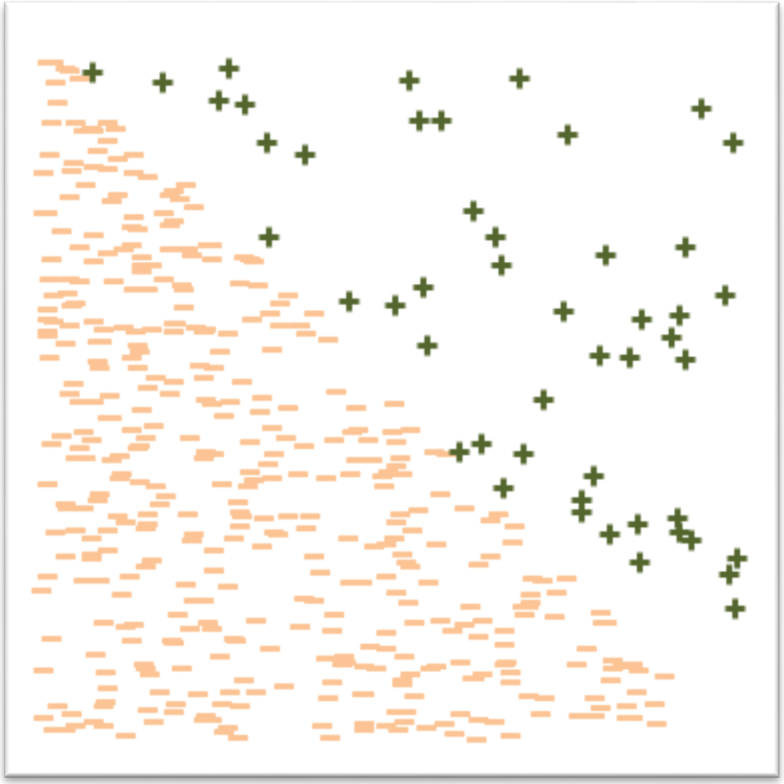
1. Search process guided by global error rates.
2. Classification rules over the positive class are highly specialized.
3. Classifiers tend to ignore small classes concentrating on classifying large ones accurately





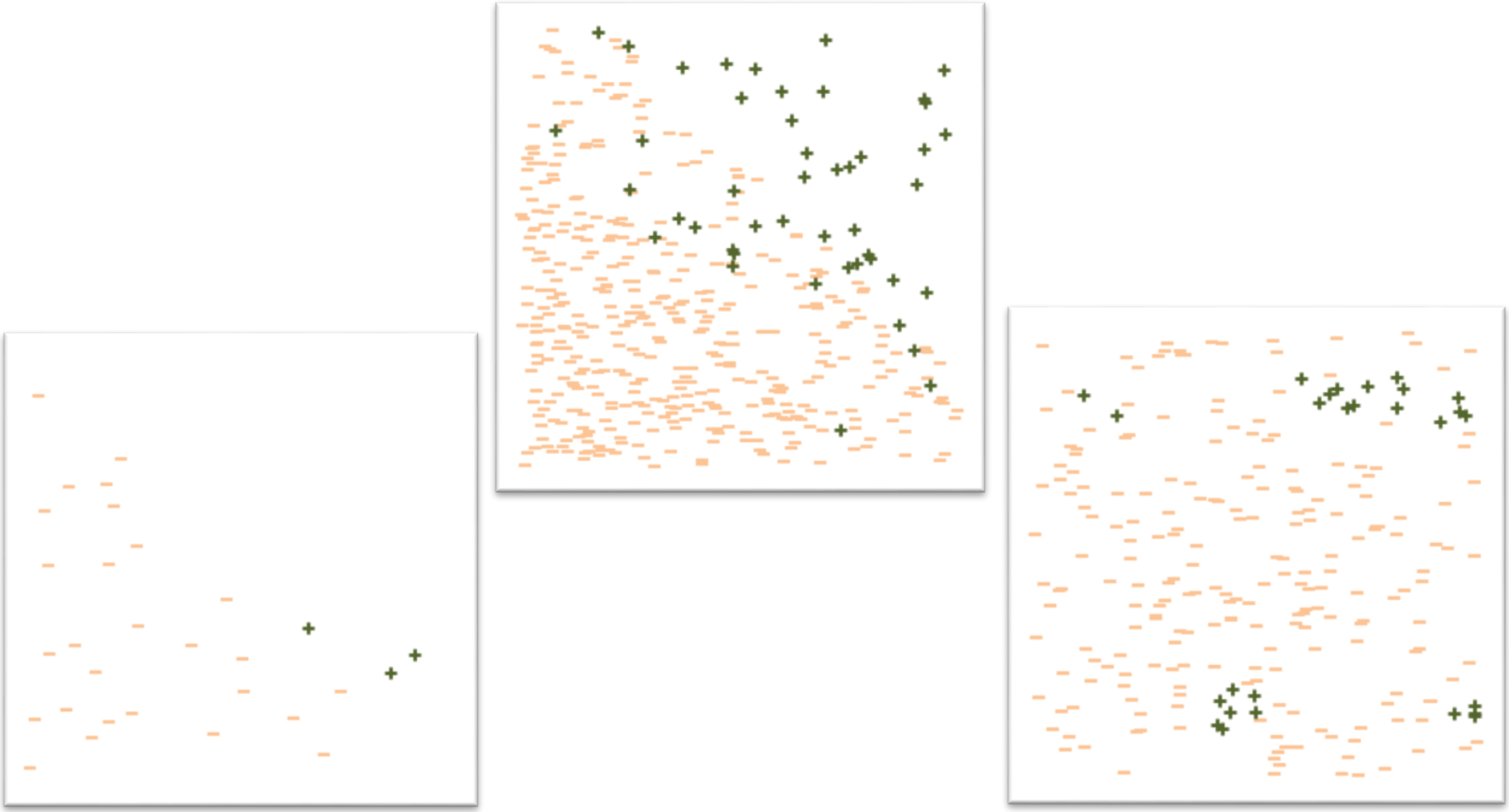
Why learning from imbalanced data-sets might be difficult?

- Skewed class distribution:
  - Measured by the fraction between majority and minority samples
  - Imbalance ratio (IR)
- **Intrinsic Data Characteristics**
  - Not only imbalance hinders classification performance
  - IR ≈ 9



V. López, A. Fernandez, S. García, V. Palade, F. Herrera, An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. Information Sciences 250 (2013) 113-141

Why learning from imbalanced data-sets might be difficult?



V. López, A. Fernandez, S. García, V. Palade, F. Herrera, An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. Information Sciences 250 (2013) 113-141

Contents

I. Introduction to imbalanced data sets

II. Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

III. Class imbalance: Data sets, implementations, …

IV. Class imbalance: Trends and final comments

Contents

I. Introduction to imbalanced data sets

II. Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

III. Class imbalance: Data sets, implementations, …

IV. Class imbalance: Trends and final comments

Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

Resampling the original training set

Cost Modifying: Cost-sensitive learning

Ensembles to address class imbalance

Introduction to Imbalanced Data Sets

Some recent applications

- **Significance of the topic in recent applications**

  - **Tan, Shing Chiang; Watada, Junzo; Ibrahim, Zuwairie; et ál.; Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects. IEEE Transactions on Neural Networks and Learning Systems 26 (5): 933-950 (MAY 2015)**
  - **Danenas, Paulius; Garsva, Gintautas; Selection of Support Vector Machines based classifiers for credit risk domain Experty Systems with Applications 42 (6) : 3194-3204 (APR 2015)**
  - **Liu, Nan; Koh, Zhi Xiong; Chua, Eric Chern-Pin; et ál..; Risk Scoring for Prediction of Acute Cardiac Complications from Imbalanced Clinical Data. IEEE Journal of Biomedical and Health Informatics 18 (6) : 1894-1902 (NOV 2014)**

Introduction to Imbalanced Data Sets

Some recent applications

- **Significance of the topic in recent applications**


  - **Radtke, Paulo V. W.; Granger, Eric; Sabourin, Robert; et ál..; Skew-sensitive boolean combination for adaptive ensembles - An application to face recognition in video surveillance Information Fusion 20: 31-48 (NOV 2014)**
  - **Yu, Hualong; Ni, Jun; An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data IEEE-ACM Transactions on Computational Biology and Bioinformatics 11(4) : 657-666 (AUG 2014)**
  - **Wang, Kung-Jeng; Makond, Bunjira; Chen, Kun-Huang; et ál.; A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Computing 20: 15-24 (JUL 2014)**
  - **B. Krawczyk, M. Galar, L. Jelen, F. Herrera. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing 38 (2016) 714-726.**

Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

Resampling the original training set

Cost Modifying: Cost-sensitive learning

Ensembles to address class imbalance

Introduction to Imbalanced Data Sets

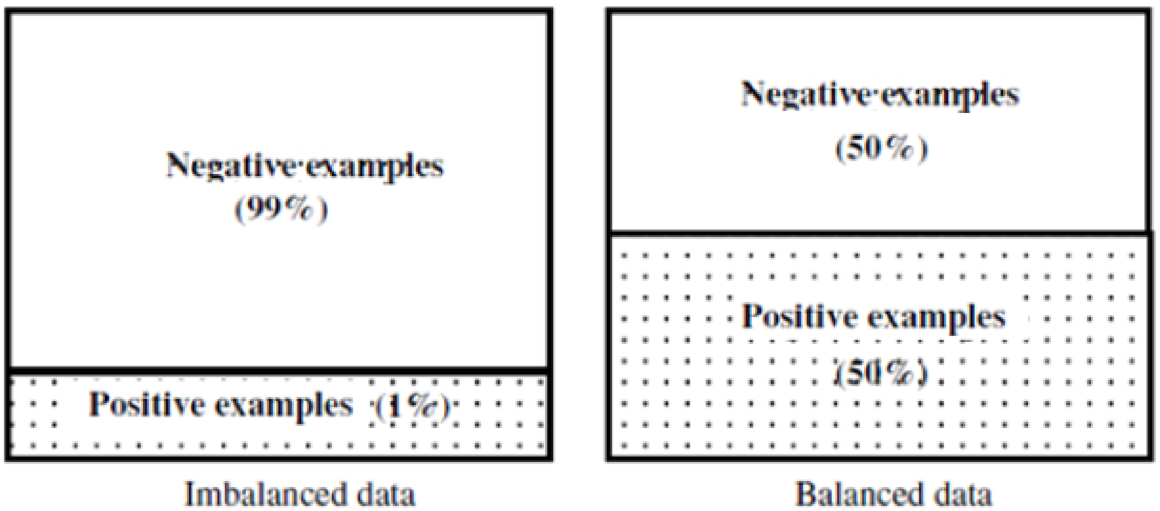


Fig. 1. Imbalanced and balanced data sets.

biased towards the majority class

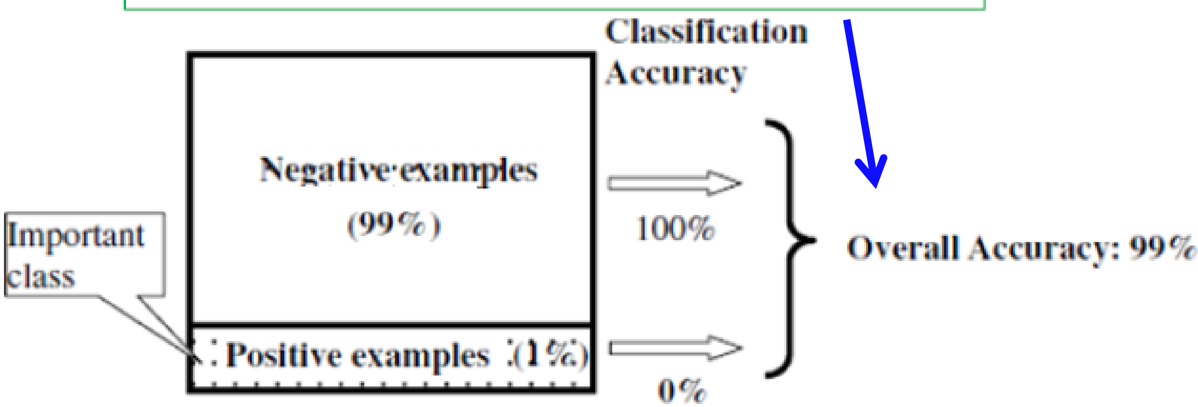


Fig. 2. The illustration of class imbalance problems.

Imbalanced classes problem: standard learners are often biased towards the majority class.

We need to change the way to evaluate a model performance!

Introduction to Imbalanced Data Sets

How can we evaluate an algorithm in imbalanced domains?

Confusion matrix for a two-class problem

	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Classical evaluation:

Error Rate: (FP + FN)/N
Accuracy Rate: (TP + TN) /N

It doesn't take into account the False Negative Rate, which is very important in imbalanced problems

Introduction to Imbalanced Data Sets

Imbalanced evaluation based on the geometric mean:

Positive true ratio: $a^+ = TP/(TP+FN)$
Negative true ratio: $a^- = TN / (FP+TN)$

Evaluation function: **True ratio**

$$g = \sqrt{(a^+ \cdot a^-)}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

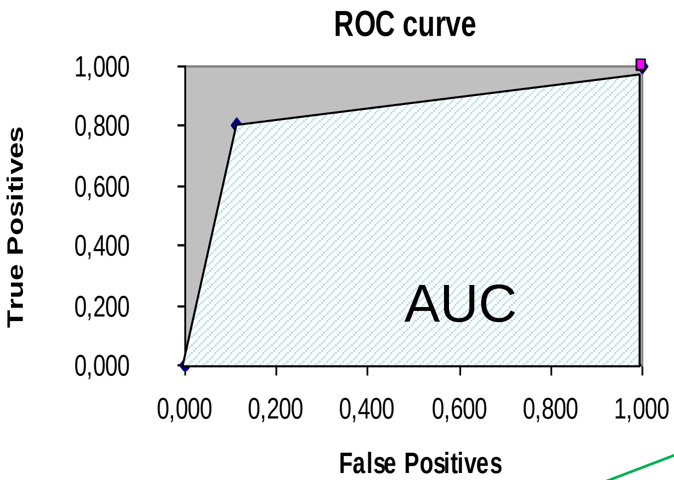
$$Specificity = \frac{TN}{TN + FP}$$

Precision = TP/(TP+FP)
Recall = TP/(TP+FN)
F-measure: (2 x precision x recall) / (recall + precision)

R. Barandela, J.S. Sánchez, V. García, E. Rangel. Strategies for learning in class imbalance problems. Pattern Recognition 36:3 (2003) 849-851

Introduction to Imbalanced Data Sets

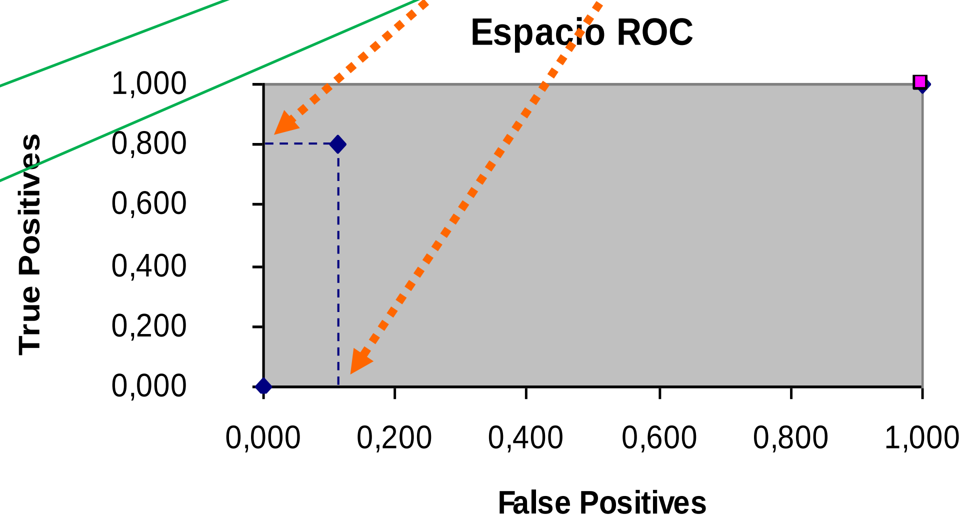
AUC: Area under ROC curve. Scalar quantity widely used for estimating classifiers performance.



Real

Pred		PP	NP
	PC	0,8	0,121
	NC	0,2	0,879

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$



Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

Resampling the original training set

Cost Modifying: Cost-sensitive learning

Ensembles to address class imbalance

Introduction to Imbalanced Data Sets

Data level vs Algorithm Level

Strategies to deal with imbalanced data sets

Motivation

- **Over-Sampling**
  - Random
  - Focused

Retain influential examples

Balance the training set

- **Under-Sampling**
  - Random
  - Focused

Remove noisy instances in the decision boundaries

Reduce the training set

- Cost Modifying (cost-sensitive)
- Algorithm-level approaches: A commont strategy to deal with the class imbalance is to choose an appropriate inductive bias.
- Boosting approaches: ensemble learning, AdaBoost, …

Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

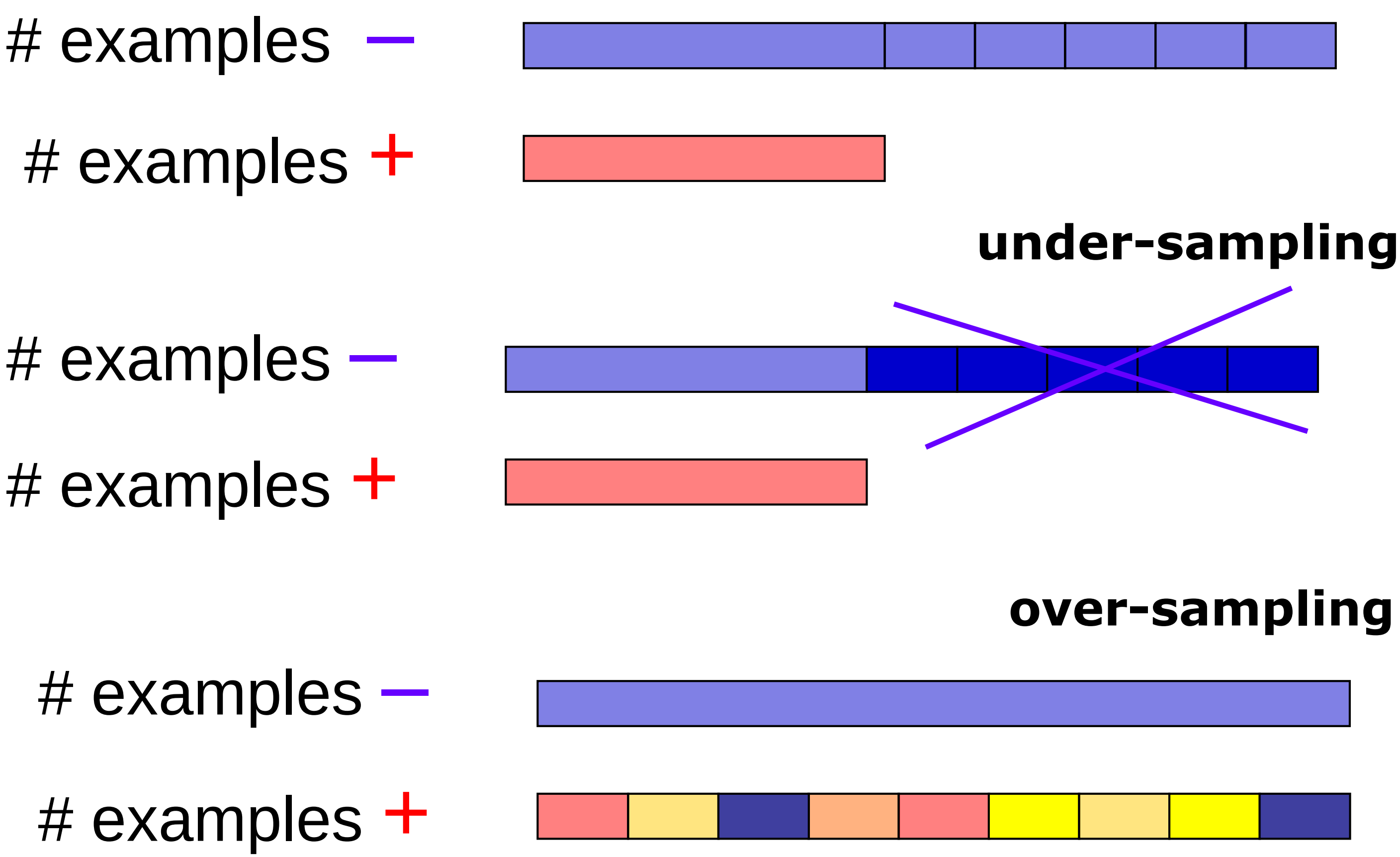
Resampling the original training set

Cost Modifying: Cost-sensitive learning

Ensembles to address class imbalance

Resampling the original data sets

Undersampling vs oversampling



Resampling the original data sets

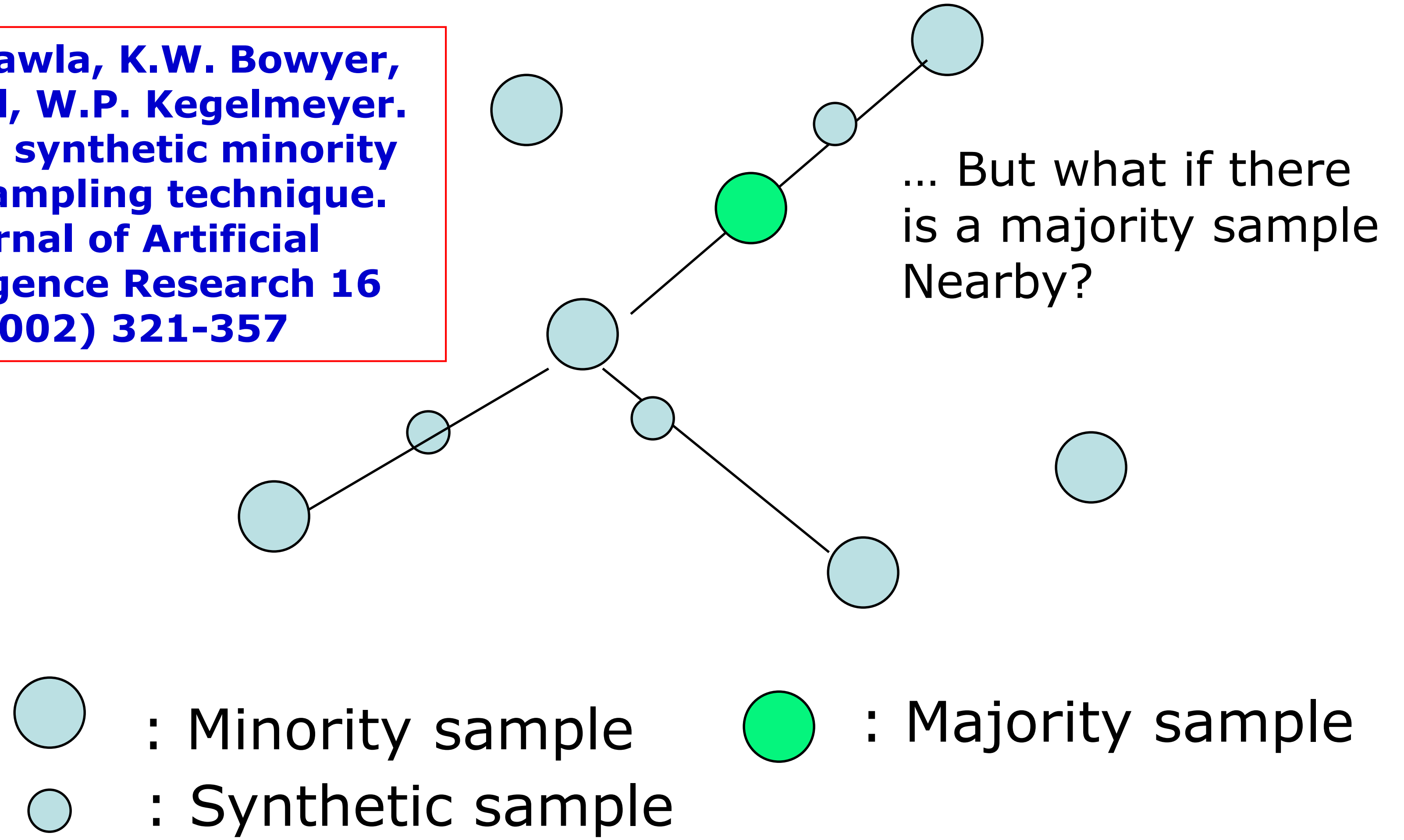
Oversampling: Replicating examples

SMOTE: Instead of replicating, let us invent some new instances.

Resampling the original data sets

Oversampling: State-of-the-art algorithm, SMOTE

N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 (2002) 321-357



Resampling the original data sets

Oversampling method:

SMOTE

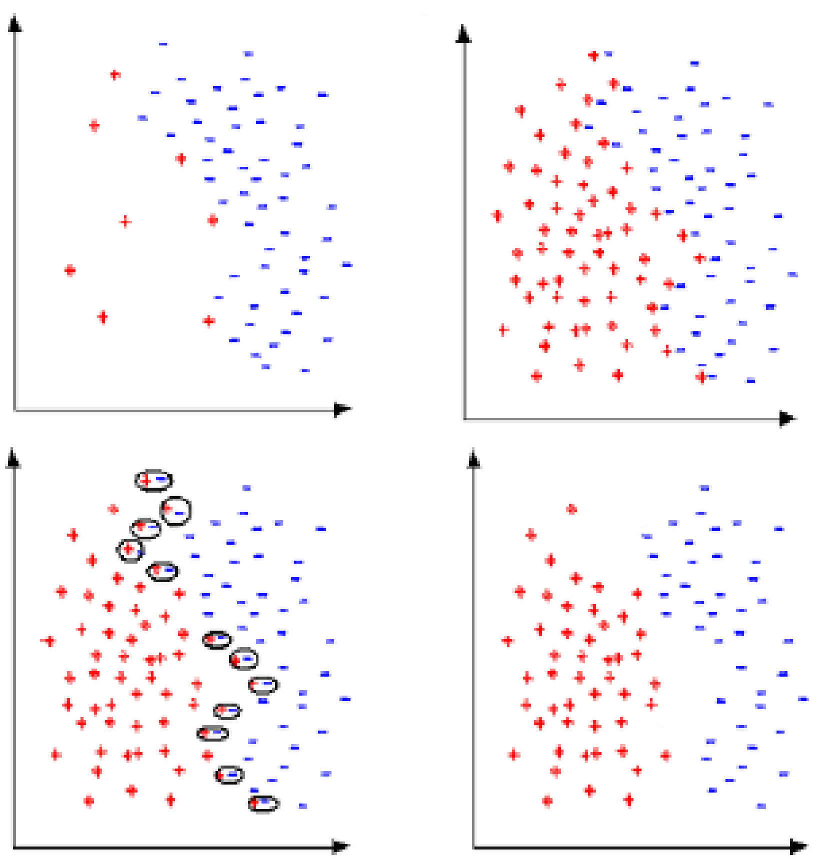
Example

of a run

Resampling the original data sets

SMOTE hybridization: SMOTE + Tomek links

Figure: SMOTE+TomekLink



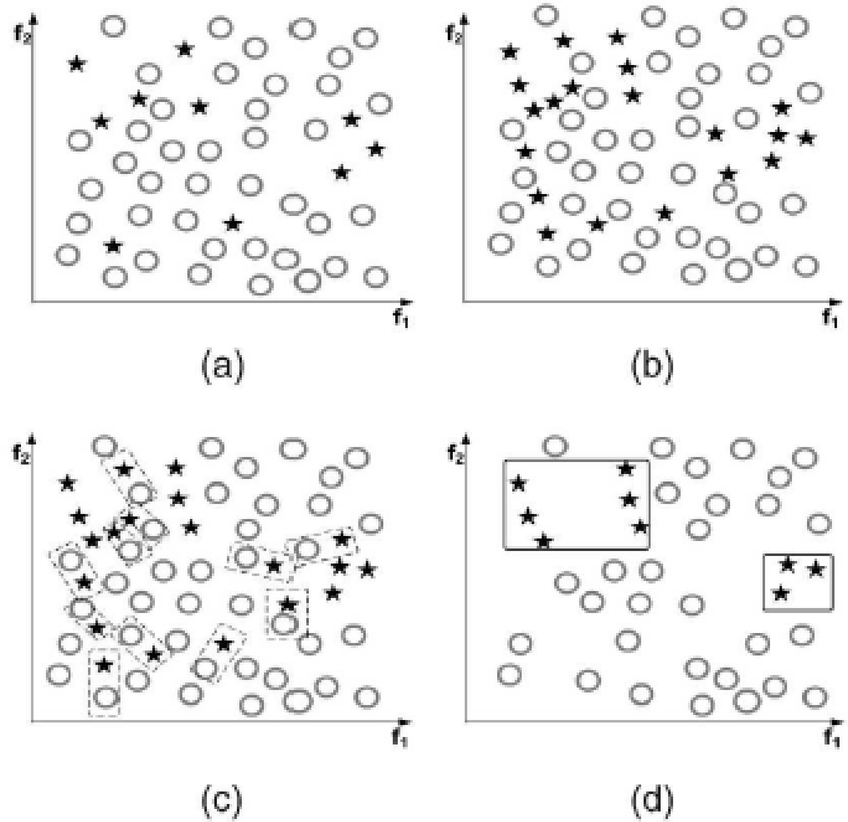


Figure 17: (a) Original data-set distribution. (b) Post-SMOTE data-set. (c) The identified Tomek Links. (d) The data-set after removing Tomek links

Resampling the original data sets

SMOTE hybridization: SMOTE + ENN

- ENN removes any example whose class label differs from the class of at least two of their neighbors
- ENN remove more examples than the Tomek links does
- ENN remove examples from both classes

Resampling the original data sets

SMOTE and hybridization: Analysis

Table 6: Performance ranking for original and balanced data sets for pruned decision trees.

Data set	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°
Pima	Smt	RdOvr	Smt+Tmk	Smt+ENN	Tmk	NCL	Original	RdUdr	CNN+Tmk	CNN*	OSS*
German	RdOvr	Smt+Tmk	Smt+ENN	Smt	RdUdr	CNN	CNN+Tmk*	OSS*	Original*	Tmk*	NCL*
Post-operative	RdOvr	Smt+ENN	Smt	Original	CNN	RdUdr	CNN+Tmk	OSS*	Tmk*	NCL*	Smt+Tmk*
Haberman	Smt+ENN	Smt+Tmk	Smt	RdOvr	NCL	RdUdr	Tmk	OSS*	CNN*	Original*	CNN+Tmk*
Splice-ie	RdOvr	Original	Tmk	Smt	CNN	NCL	Smt+Tmk	Smt+ENN*	CNN+Tmk*	RdUdr*	OSS*
Splice-ei	Smt	Smt+Tmk	Smt+ENN	CNN+Tmk	OSS	RdOvr	Tmk	CNN	NCL	Original	RdUdr
Vehicle	RdOvr	Smt	Smt+Tmk	OSS	CNN	Original	CNN+Tmk	Tmk	NCL*	Smt+ENN*	RdUdr*
Letter-vowel	Smt+ENN	Smt+Tmk	Smt	RdOvr	Tmk*	NCL*	Original*	CNN*	CNN+Tmk*	RdUdr*	OSS*
New-thyroid	Smt+ENN	Smt+Tmk	Smt	RdOvr	RdUdr	CNN	Original	Tmk	CNN+Tmk	NCL	OSS
E.Coli	Smt+Tmk	Smt	Smt+ENN	RdOvr	NCL	Tmk	RdUdr	Original	OSS	CNN+Tmk*	CNN*
Satimage	Smt+ENN	Smt	Smt+Tmk	RdOvr	NCL	Tmk	Original*	OSS*	CNN+Tmk*	RdUdr*	CNN*
Flag	RdOvr	Smt+ENN	Smt+Tmk	CNN+Tmk	Smt	RdUdr	CNN*	OSS*	Tmk*	Original*	NCL*
Glass	Smt+ENN	RdOvr	NCL	Smt	Smt+Tmk	Original	Tmk	RdUdr	CNN+Tmk*	OSS*	CNN*
Letter-a	Smt+Tmk	Smt+ENN	Smt	RdOvr	OSS	Original	Tmk	CNN+Tmk	NCL	CNN	RdUdr*
Nursery	RdOvr	Tmk	Original	NCL	CNN*	OSS*	Smt+Tmk*	Smt*	CNN+Tmk*	Smt+ENN*	RdUdr*

G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29

Resampling the original data sets

Other SMOTE hybridizations

Safe_Level_SMOTE: C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09). LNAI 5476, Springer-Verlag 2005, Bangkok (Thailand, 2009) 475-482

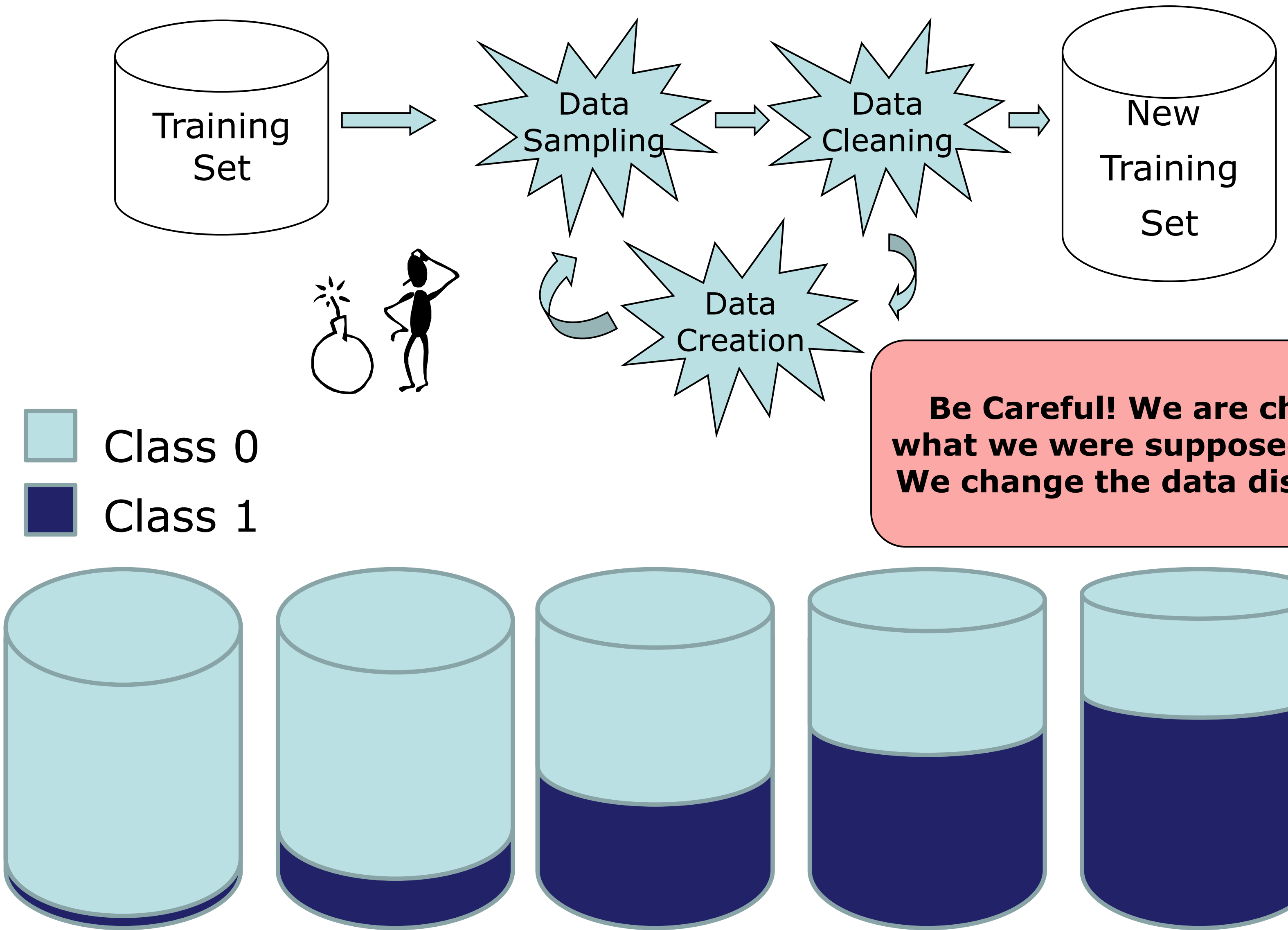
Borderline_SMOTE: H. Han, W.Y. Wang, B.H. Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing (ICIC'05). Lecture Notes in Computer Science 3644, Springer-Verlag 2005, Hefei (China, 2005) 878-887

SMOTE_LLE: J. Wang, M. Xu, H. Wang, J. Zhang. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. IEEE 8th International Conference on Signal Processing, 2006.

LN-SMOTE: T. Maciejewski and J. Stefanowski. Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data. IEEE SSCI , Paris, CIDM , 2011.

SMOTE-RSB: E. Ramentol, Y. Caballero, R. Bello, F. Herrera, SMOTE-RSB*: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory. *Knowledge and Information Systems 33:2 (2012) 245-265.*

Resampling the original data sets
Final comments



Be Careful! We are changing what we were supposed to learn! We change the data distribution!

Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

Resampling the original training set

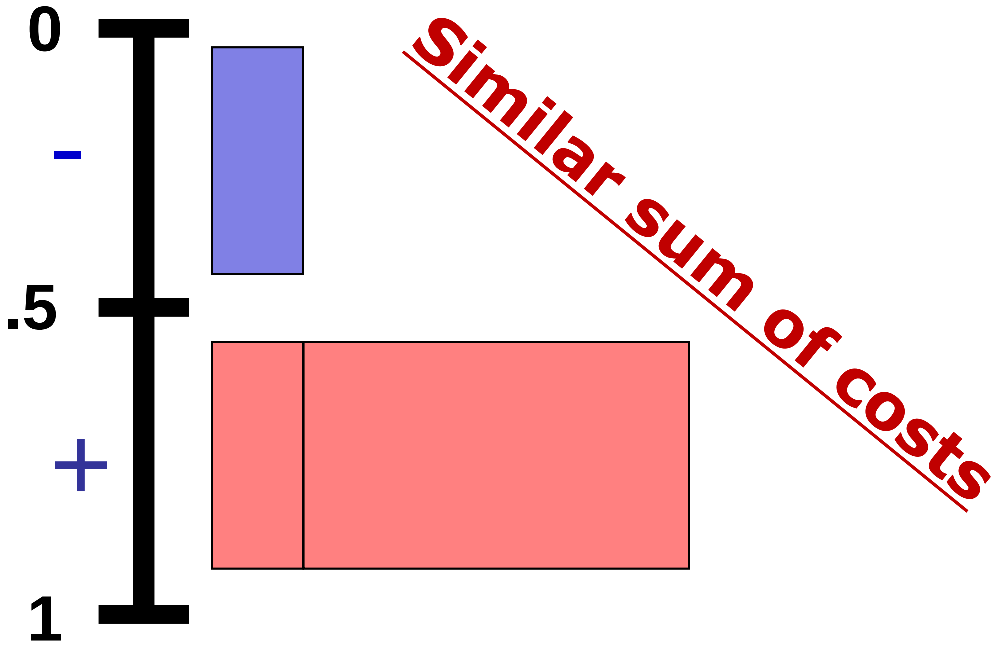
Cost Modifying: Cost-sensitive learning

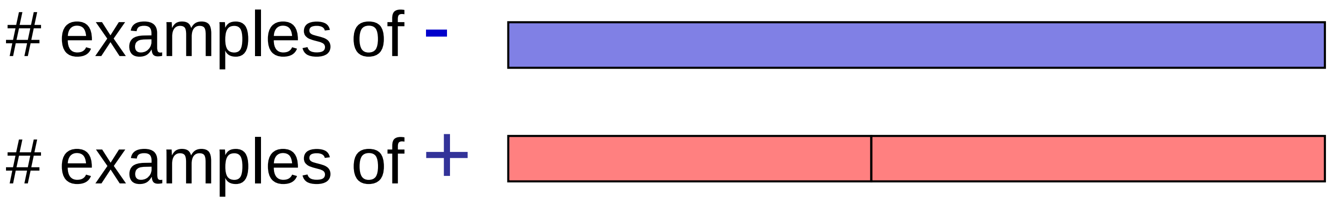
Ensembles to address class imbalance

Cost-sensitive learning

Cost modification consists of weighting errors made on examples of the minority class higher than those made on examples of the majority class in the calculation of the training error.

- Over Sampling
  - Random
  - Focused
- Under Sampling
  - Random
  - Focused
- **Cost Modifying**





Cost-sensitive learning

Results and Statistical Analysis

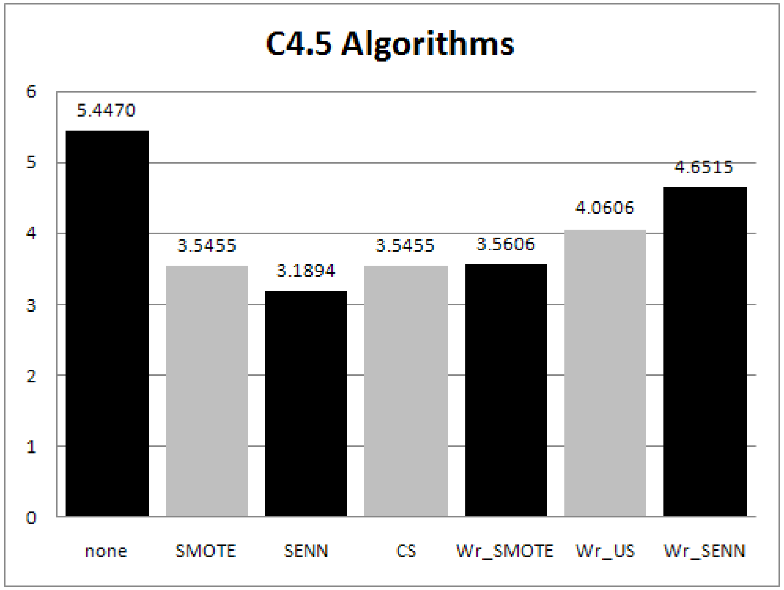
- **Case of Study: C4.5**
- **Similar results and conclusions for the remaining classification paradigms**

Algorithm	AUC_{tr}	AUC_{tst}
C45	0.8774 ± 0.0392	0.7902 ± 0.0804
C45 SMOTE	0.9606 ± 0.0142	0.8324 ± 0.0728
C45 SENN	0.9471 ± 0.0154	$\mathbf{0.8390 \pm 0.0772}$
C45CS	0.9679 ± 0.0103	0.8294 ± 0.0758
C45 Wr_SMOTE	0.9679 ± 0.0103	0.8296 ± 0.0763
C45 Wr_US	0.9635 ± 0.0139	0.8245 ± 0.0760
C45 Wr_SENN	0.9083 ± 0.0377	0.8145 ± 0.0712

V. López, A. Fernandez, J. G. Moreno-Torres, F. Herrera, **Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics**. *Expert Systems with Applications 39:7 (2012) 6585-6608.*

Cost-sensitive learning

Results and Statistical Analysis



- Rankings obtained by Friedman test for the different approaches of C4.5.
- Shaffer test as post-hoc to detect statistical differences ($\alpha = 0.05$)

C4.5	none	SMOTE	SENN	CS	Wr_SMOTE	Wr_US	Wr_SENN
none	x	(6.404E-6)	(4.058E-8)	(6.404E-6)	(7.904E-6)	(.00341)	=(.37846)
SMOTE	+(6.404E-6)	x	=(1.0)	=(1.0)	=(1.0)	=(1.0)	+(.04903)
SENN	+(4.058E-8)	=(1.0)	x	=(1.0)	=(1.0)	=(.22569)	+(.00152)
CS	+(6.404E-6)	=(1.0)	=(1.0)	x	=(1.0)	=(1.0)	+(.04903)
Wr_SMOTE	+(7.904E-6)	=(1.0)	=(1.0)	=(1.0)	x	=(1.0)	+(.04903)
Wr_US	+(.00341)	=(1.0)	=(.22569)	=(1.0)	=(1.0)	x	=(1.0)
Wr_SENN	=(.37846)	-(.04903)	-(.00152)	-(.04903)	-(.04903)	=(1.0)	x

Cost-sensitive learning
Final comments

- **Preprocessing and cost-sensitive learning improve the base classifier.**
- **No differences among the different preprocessing techniques.**
- **Both preprocessing and cost-sensitive learning are good and equivalent approaches to address the imbalance problem.**
- **In most cases, the preliminary versions of hybridization techniques do not show a good behavior in contrast to standard preprocessing and cost sensitive.**

Some authors claim: **"Cost-Adjusting is slightly more effective than random or directed over- or under-sampling although all approaches are helpful, and directed oversampling is close to cost-adjusting". Our study shows similar results.**

V. López, A. Fernandez, J. G. Moreno-Torres, F. Herrera, **Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics**. *Expert Systems with Applications 39:7 (2012) 6585-6608.*

Introduction to Imbalanced Data Sets

Some recent applications

How can we evaluate an algorithm in imbalanced domains?

Strategies to deal with imbalanced data sets

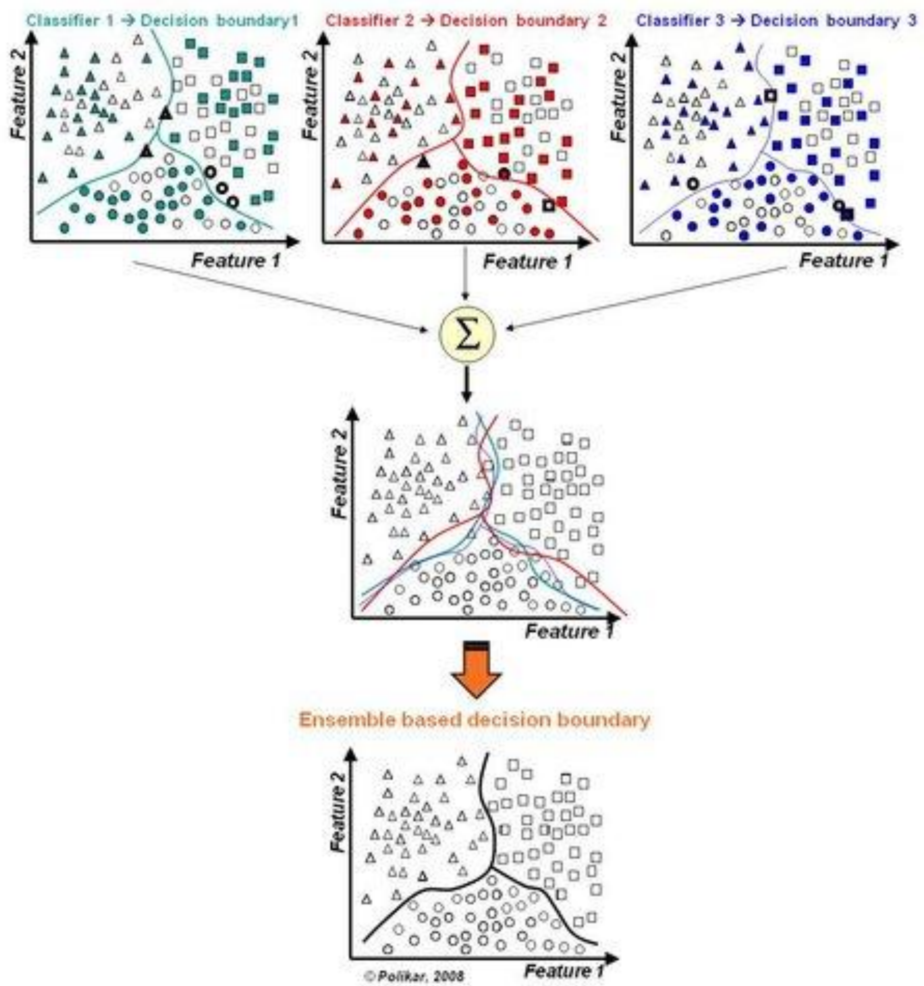
Resampling the original training set

Cost Modifying: Cost-sensitive learning

Ensembles to address class imbalance

Emsembles to address class imbalance

Ensemble-based classifiers try to improve the performance of single classifiers by inducing several classifiers and combining them to obtain a new classifier that outperforms every one of them.
Hence, the basic idea is to construct several classifiers from the original data and then aggregate their predictions when unknown instances are presented.
This idea follows human natural behavior which tend to seek several opinions before making any important decision.



Emsembles to address class imbalance

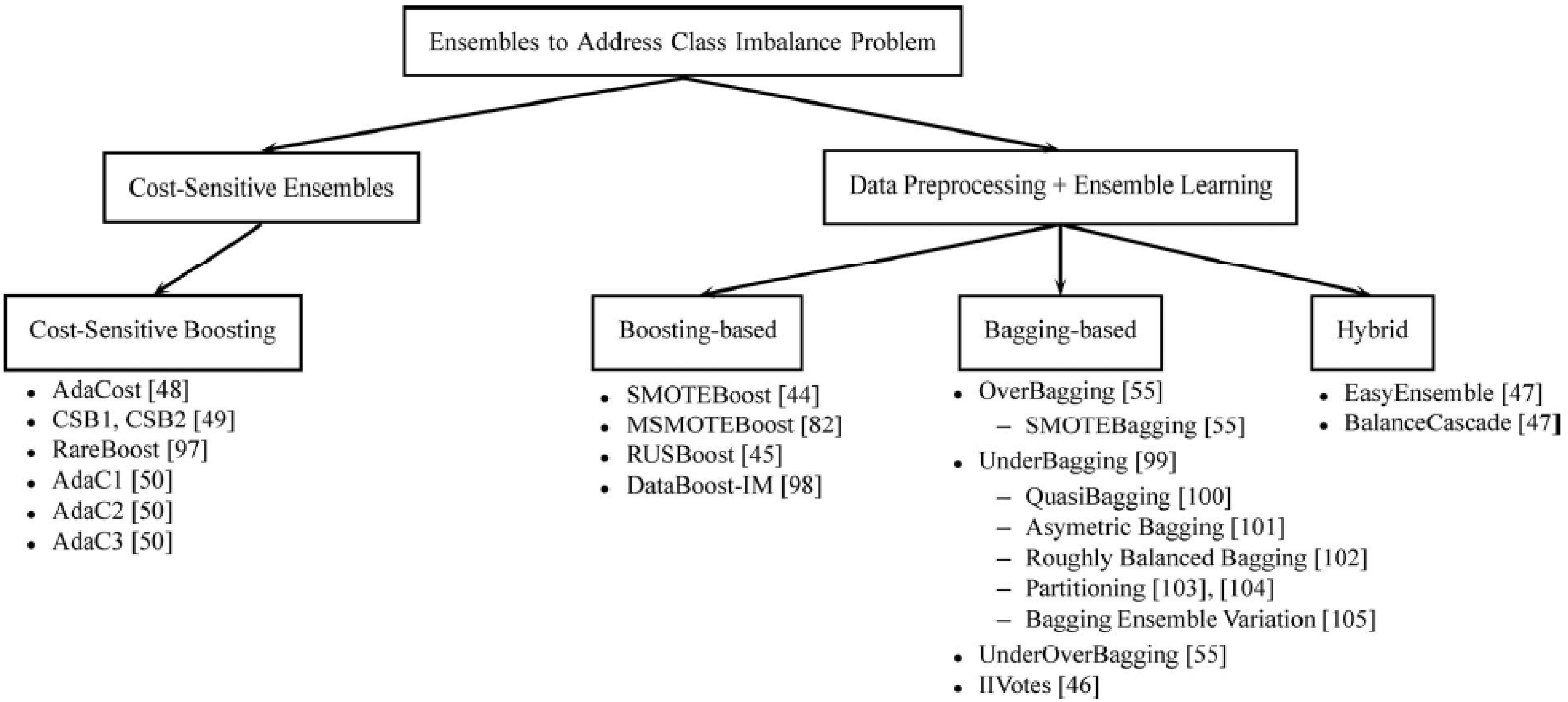


Fig. 3. Proposed taxonomy for ensembles to address the class imbalance problem.

M. Galar, A. Fernández, F. E. Barrenechea, H. Bustince, F. Herrera. **A Review on Ensembles for Class Imbalance Problem: Bagging, Boosting and Hybrid Based Approaches**. IEEE TSMC-Par C **42:4 (2012) 463-484**

Emsembles to address class imbalance

TABLE XV
REPRESENTATIVE METHODS SELECTED FOR EACH FAMILY

Family	Abbr.	Method
Non-ensembles	SMT	SMOTE
Classic	M14	AdaBoost.M2 ($T = 40$)
Cost-sensitive	C24	AdaC2 ($T = 40$)
Boosting-based	RUS1	RUSBoost ($T = 10$)
Bagging-based	SBAG4	SMOTEBagging ($T = 40$)
Hybrids	EASY	EasyEnsemble

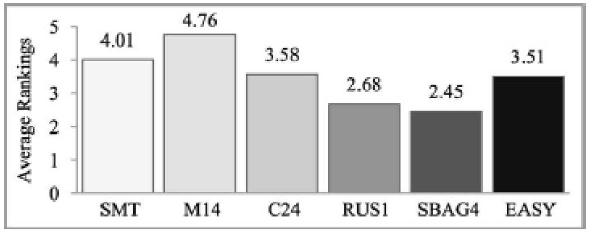


Fig. 9. Average rankings of the representatives of each family.

TABLE XVI
HOLM TABLE FOR BEST INTERFAMILY ANALYSIS

i	Algorithm (Rank)	Z	p-value	Holm	Hypothesis ($\alpha = 0.05$)
5	M14 (4.76)	5.78350	0.00000	0.01	**Rejected for SBAG4**
4	SMT (4.01)	3.90315	0.00009	0.0125	**Rejected for SBAG4**
3	C24 (3.58)	2.82052	0.00479	0.01667	**Rejected for SBAG4**
2	EASY (3.51)	2.64958	0.00806	0.025	**Rejected for SBAG4**
1	RUS1 (2.68)	0.56980	0.56881	0.05	Not Rejected

Control method : SBAG4, Rank :2.45.

TABLE XVIII
SHAFFER TESTS FOR INTERFAMILY COMPARISON

	SMT	M14	C24	RUS1	SBAG4	EASY
SMT	×	=(0.24024)	=(1.0)	-(0.00858)	-(0.00095)	=(1.0)
M14	=(0.24024)	×	-(0.03047)	-(0.0)	-(0.0)	-(0.01725)
C24	=(1.0)	+(0.03047)	×	=(0.17082)	-(0.03356)	=(1.0)
RUS1	+(0.00858)	+(0.0)	=(0.17082)	×	=(1.0)	=(0.22527)
SBAG4	+(0.00095)	+(0.0)	+(0.03356)	=(1.0)	×	=(0.05641)
EASY	+(0.01725)	=(1.0)	=(1.0)	=(0.22527)	=(0.05641)	×

TABLE XVII
WILCOXON TESTS TO SHOW DIFFERENCES BETWEEN SBAG4 AND RUS1

Comparison	R^+	R^-	Hypothesis($\alpha = 0.05$)	p-value
SBAG4 vs. RUS1	527.5	462.5	Not Rejected	0.71717

R^+ are ranks for SBAG4 and R^- for RUS1.

Emsembles to address class imbalance

Our proposal:

We develop a new ensemble construction algorithm (**EUSBoost**) based on RUSBoost, one of the simplest and most accurate ensemble, combining random undersampling with Boosting algorithm.

Our methodology aims to improve the existing proposals enhancing the performance of the base classifiers by the usage of the evolutionary undersampling approach.

Besides, we promote diversity favoring the usage of different subsets of majority class instances to train each base classifier.

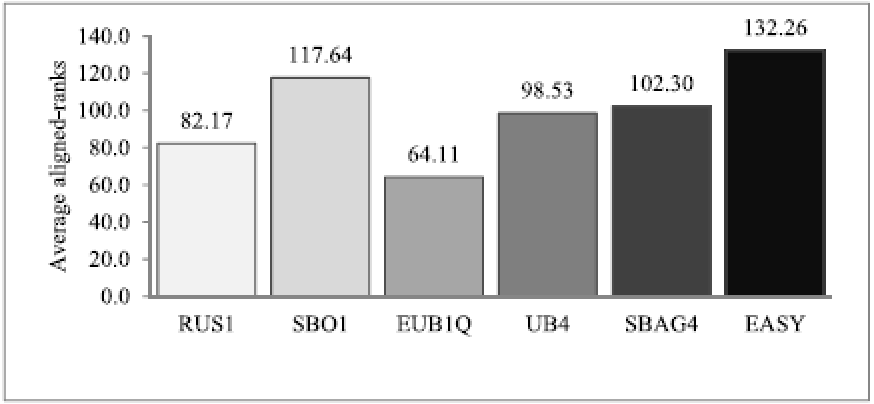


Figure: Average aligned-ranks of the comparison between EUSBoost and the state-of-the-art ensemble methods.

M. Galar, A. Fernandez, E. Barrenechea, F. Herrera, **EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-sets by Evolutionary Undersampling**. *Pattern Recognition 46:12 (2013) 3460–3471*

Emsembles to address class imbalance
Final comments

- Ensemble-based algorithms are worthwhile, improving the results obtained by using data preprocessing techniques and training a single classifier.

- The use of more classifiers make them more complex, but this growth is justified by the better results that can be assessed.

- We have to remark the good performance of approaches such as RUSBoost or SmoteBagging, which despite of being simple approaches, achieve higher performance than many other more complex algorithms.

- We have shown the positive synergy between sampling techniques (e.g., undersampling or SMOTE) and Bagging ensemble learning algorithm.

Contents

Inteligencia de Negocio

TEMA 7. Modelos Avanzados de Minería de Datos

1. Clases no balanceadas/equilibradas

2. **Características intrínsecas de los datos en clasificación**

3. Detección de anomalías

4. Problemas no estándar de clasificación: MIL, MLL, …

5. Análisis de Sentimientos

6. Deep Learning

Why is difficult to learn in imbalanced domains?

- **Preprocessing and cost sensitive learning have a similar behavior.**
- **Performance can still be improved, but we must analyze in deep the nature of the imbalanced data-set problem:**
  - **Imbalance ratio is not a determinant factor**

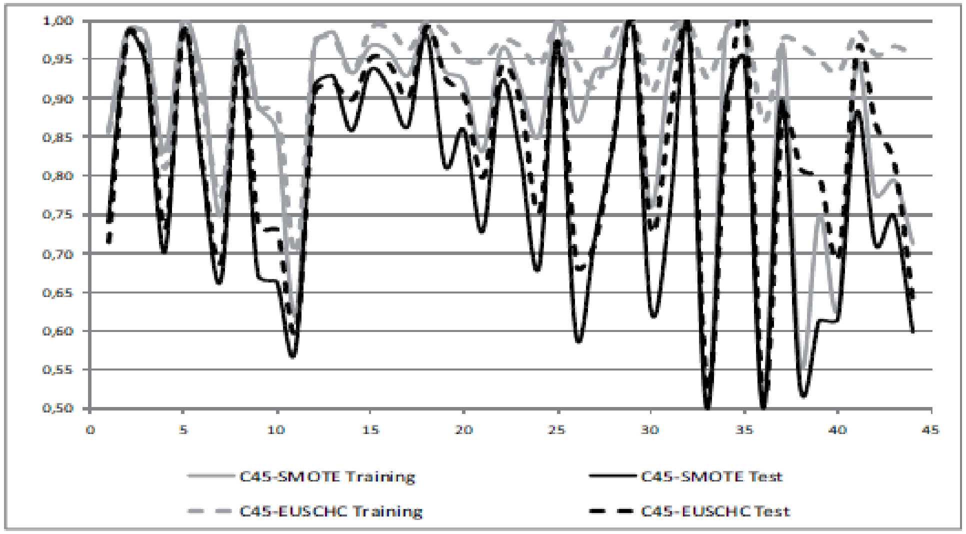
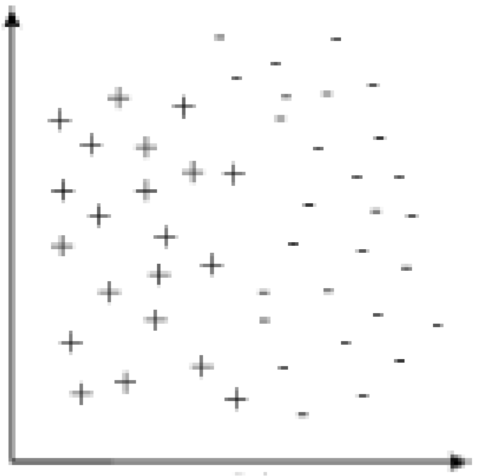


Fig. 4 C4.5 AUC in Training/Test sorted by IR

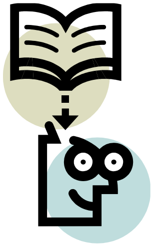
J. Luengo, A. Fernández, S. García, and F. Herrera. Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing 15 (2011) 1909-1936*, doi: 10.1007/s00500-010-0625-8.

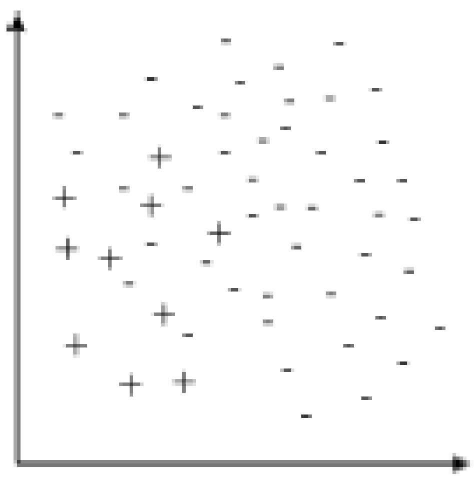
Introduction to Imbalanced Data Sets

Why is difficult to learn in imbalanced domains?



Imbalance – why is it difficult?





An easier problem

More difficult one

Some of sources of difficulties:

- **Overlapping,**
- **Small disjuncts,**
- **Lack of data,**
- **…**

Majority classes overlaps the minority class:

- **Ambiguous boundary between classes**
- **Influence of noisy examples**
- **Difficult border, …**

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

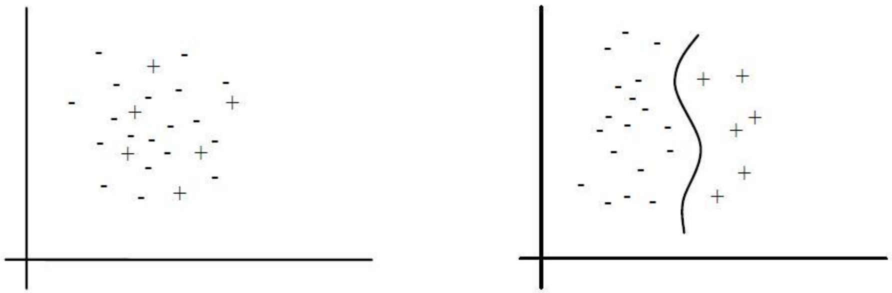
Bordeline and Noise data

Dataset shift

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

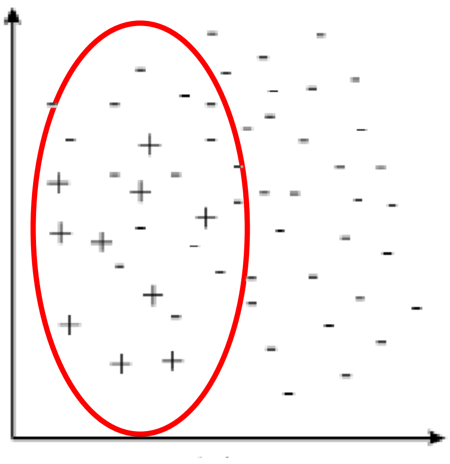
Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Class imbalance is not the only responsible of the lack in accuracy of an algorithm.



(a) (b)

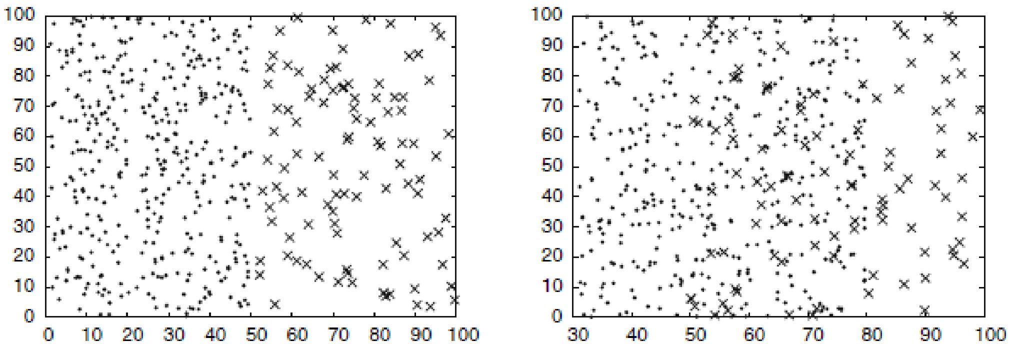
The class overlapping also influences the behaviour of the algorithms, and it is very typical in these domains.



V. García, R.A. Mollineda, J.S. Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Applic (2008) 11: 269-280

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

- **There is an interesting relationship between imbalance and class overlapping:**



(a) (b)

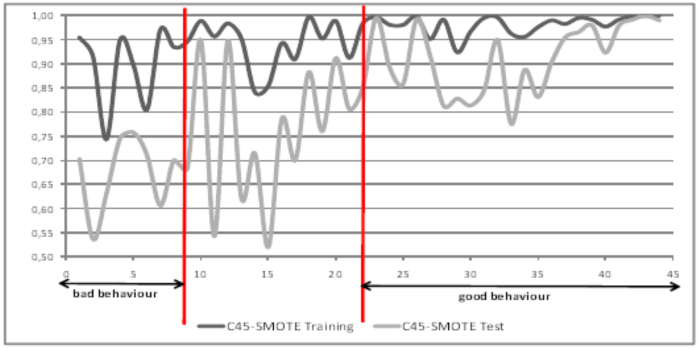


Fig. 6 C4.5 AUC with SMOTE in Training/Test sorted by F1

Two different levels of class overlapping: (a) 0% and (b) 60%

F1: maximum Fisher's discriminant ratio.

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

V. García, R.A. Mollineda, J.S. Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Applic (2008) 11: 269-280

J. Luengo, A. Fernandez, S. García, F. Herrera, Addressing Data Complexity for Imbalanced Data Sets: Analysis of SMOTE-based Oversampling and Evolutionary Undersampling. ***Soft Computing, 15 (10) 1909-1936***

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

- **There is an interesting relationship between imbalance and class overlapping:**

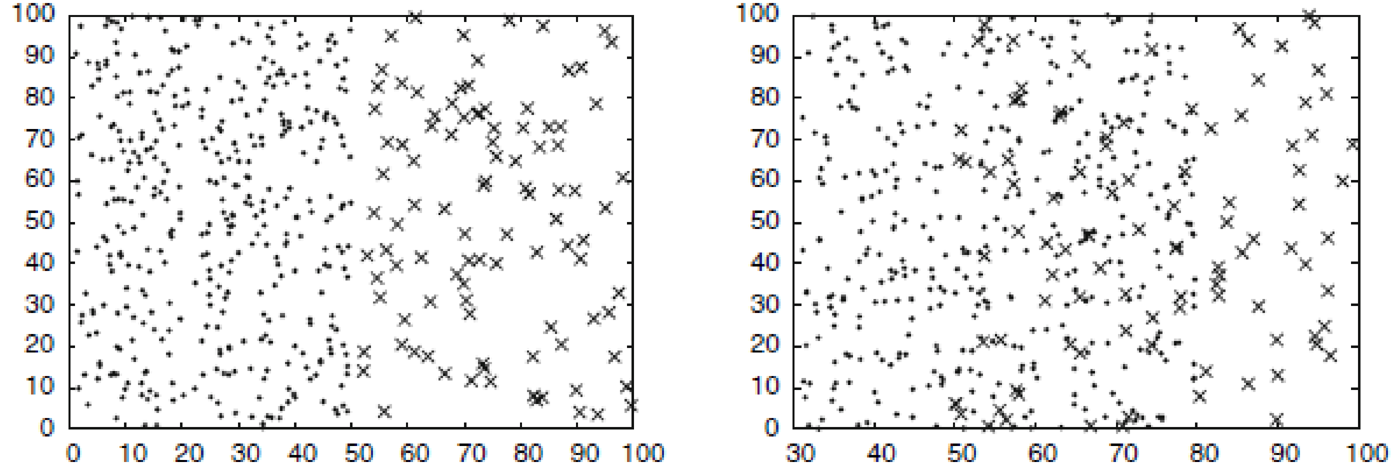
Table 13 Performance obtained by C4.5 with different degrees of overlap

Overlap Degree	TP_{rate}	TN_{rate}	AUC
0 %	1.000	1.000	1.000
20 %	.79.00	1.000	.8950
40 %	.4900	1.000	.7450
50 %	.4700	1.000	.7350
60 %	.4200	1.000	.7100
80 %	.2100	.9989	.6044
100 %	.0000	1.000	.5000

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Overlapping

V. García, R.A. Mollineda, J.S. Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Applic (2008) 11: 269-280



(a) (b)

Fig. Two different levels of class overlapping: a 0% and b 60%

Experiment I: The positive examples are defined on the X-axis in the range [50–100], while those belonging to the majority class are generated in [0–50] for 0% of class overlap, [10–60] for 20%, [20–70] for 40%, [30–80] for 60%, [40–90] for 80%, and [50–100] for 100% of overlap.
The overall imbalance ratio matches the imbalance ratio corresponding to the overlap region, what could be accepted as a common case.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

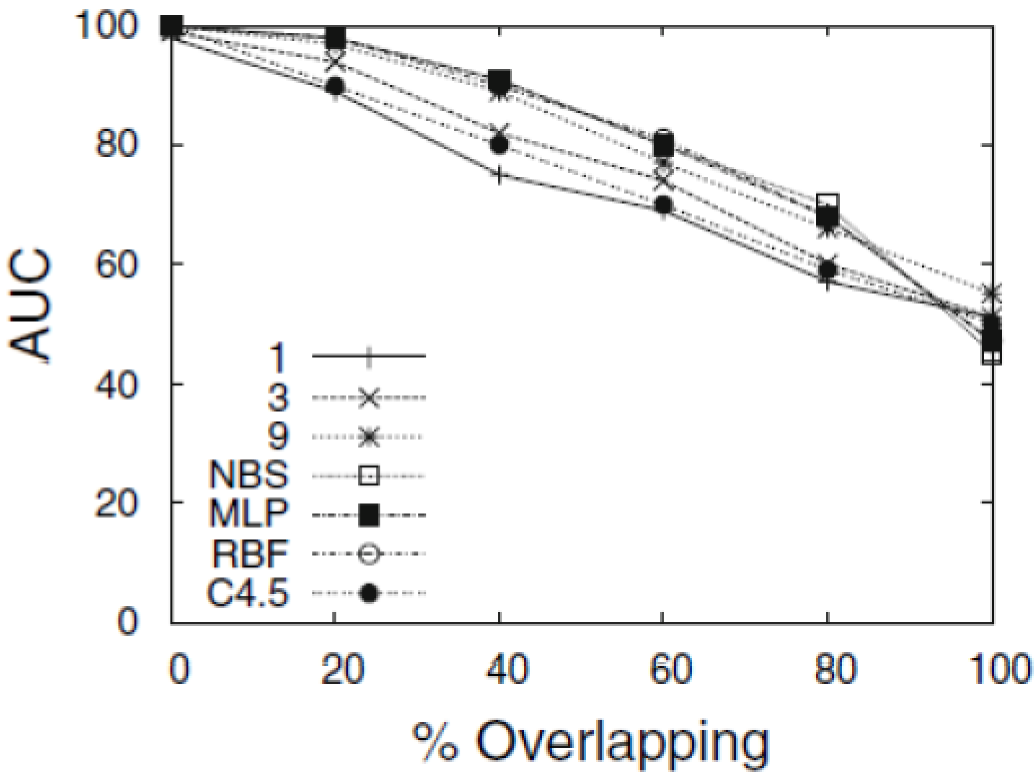


Fig. Performance metrics in k-NN rule and other learning algorithms for experiment I

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

V. García, R.A. Mollineda, J.S. Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Applic (2008) 11: 269-280

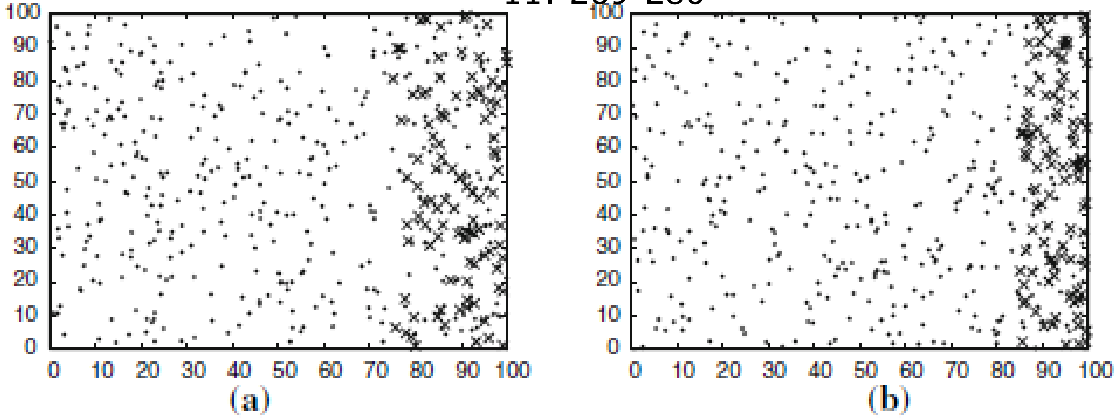


Fig. Two different cases in experiment II: [75-100] and [85-100]. For this latter case, note that in the overlap region, the majority class is under-reprsented in comparison to the minority class.

Experiment II: The second experiment has been carried out over a collection of five artificial imbalanced data sets in which the overall minority class becomes the majority in the overlap region. To this end, the 400 negative examples have been defined on the X-axis to be in the range [0–100] in all data sets, while the 100 positive cases have been generated in the ranges [75–100], [80–100], [85–100], [90–100], and [95–100]. The number of elements in the overlap region varies from no local imbalance in the first case, where both classes have the same (expected) number of patterns and density, to a critical inverse imbalance in the fifth case, where the 100 minority examples appears as majority in the overlap region along with about 20 expected negative examples.

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Overlapping

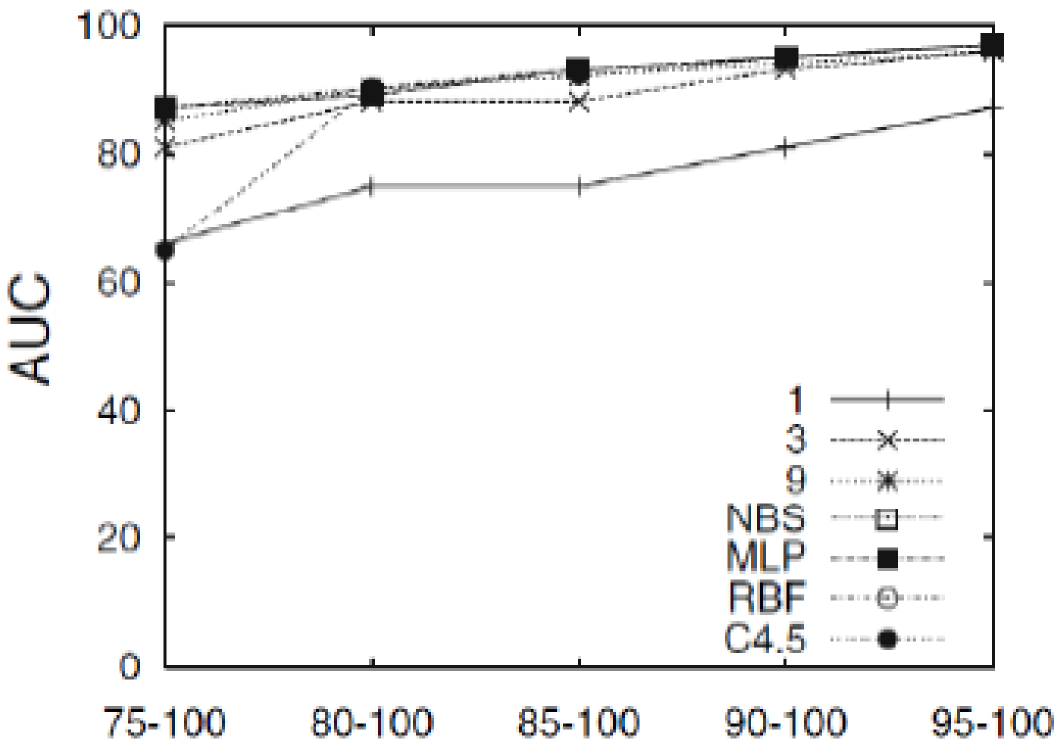
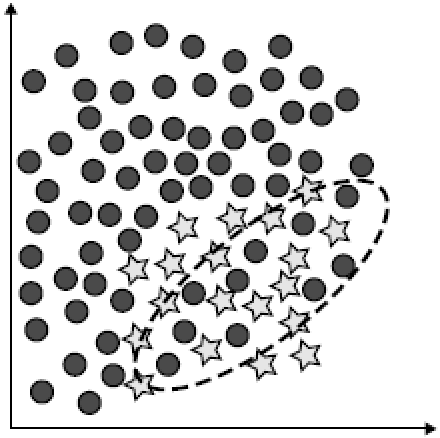


Fig. Performance metrics in k-NN rule and other learning algorithms for experiment II

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Overlapping



(a) Class overlapping

Conclusions: Results (in this paper) shov that the class more represented in overlap regions tends to be better classified by methods based on global learning, while the represented in such regions tends to be better classified by local methods.

In this sense, as the value of k of the k-NN rule increases, along with a weakening of its local nature, it was progressively approaching the behaviour of global models.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

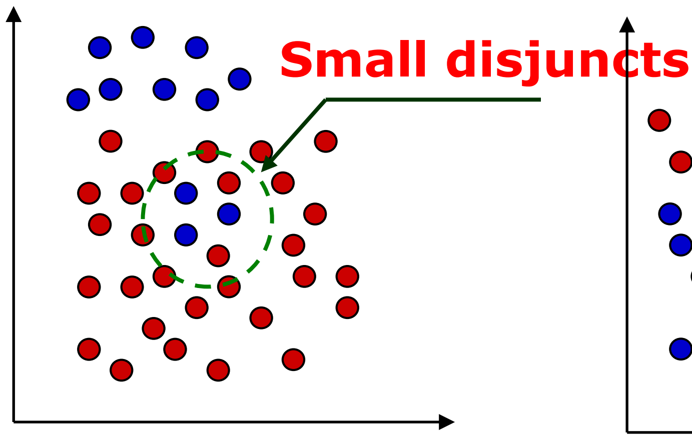
Bordeline and Noise data

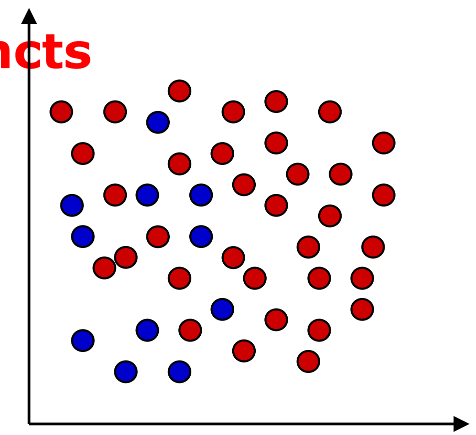
Dataset shift

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Class imbalance is not the only responsible of the lack in accuracy of an algorithm.





Class imbalances may yield small disjuncts which, in turn, will cause degradation.

Rare cases or Small disjuncts are those disjuncts in the learned classifier that cover few training examples.

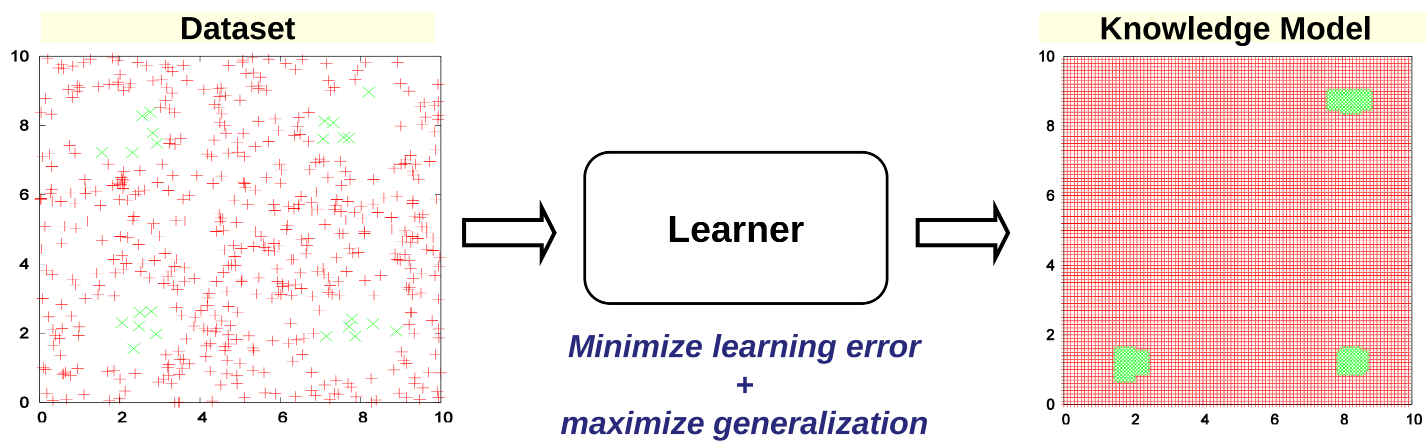
T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49
G.M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6:1 (2004) 7-19

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Why is difficult to learn in imbalanced domains?

Rare or exceptional cases correspond to small numbers of training examples in particular areas of the feature space. When learning a concept, the presence of rare cases in the domain is an important consideration. The reason why rare cases are of interest is that they cause small disjuncts to occur, which are known to be more error prone than large disjuncts.

In the real world domains, rare cases are unknown since high dimensional data cannot be visualized to reveal areas of low coverage.



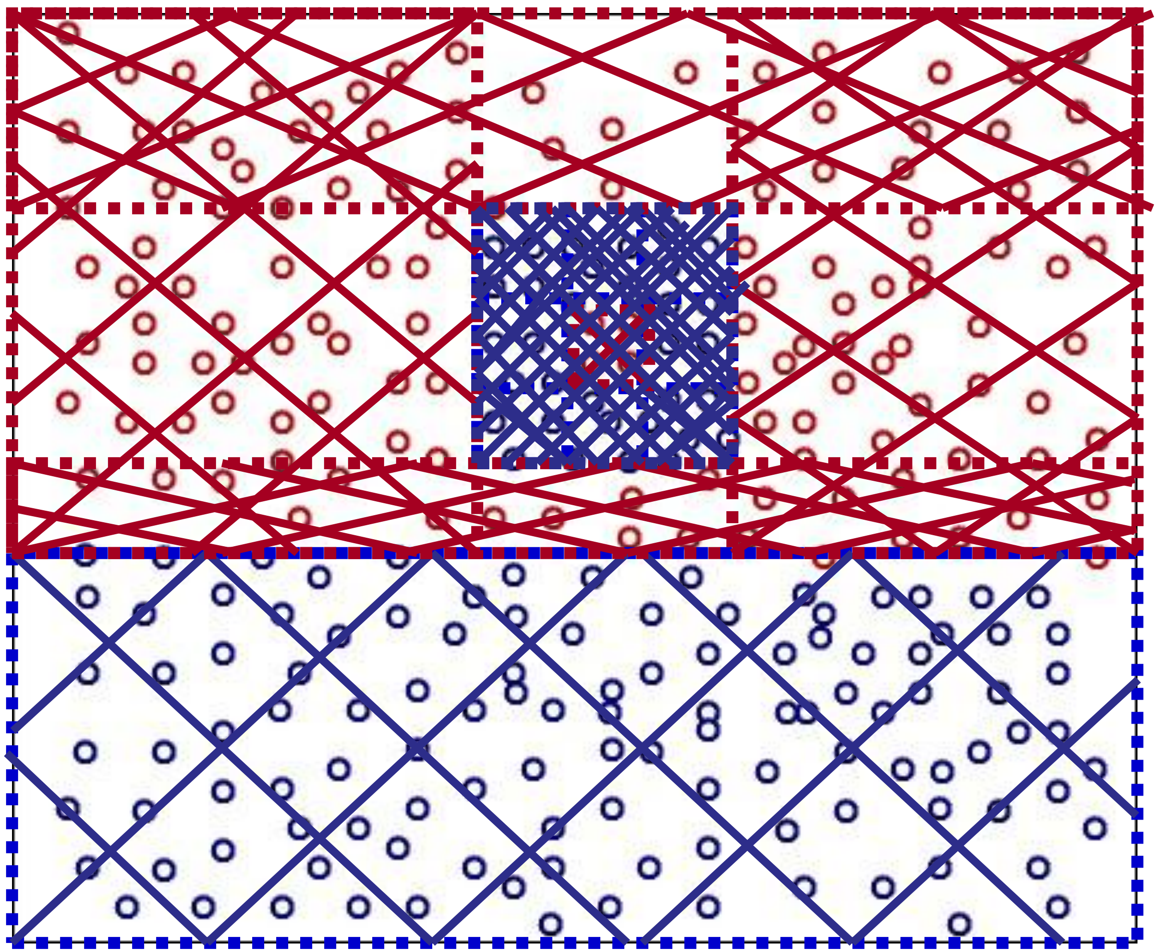
T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Rare or excepcional cases

Rare cases or Small disjunct: Focusing the problem



Small Disjunct or *Starved niche*

Again
more small disjuncts

Overgeneral Classifier

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Rare or excepcional cases

Rarity: Rare Cases versus Rare Classes

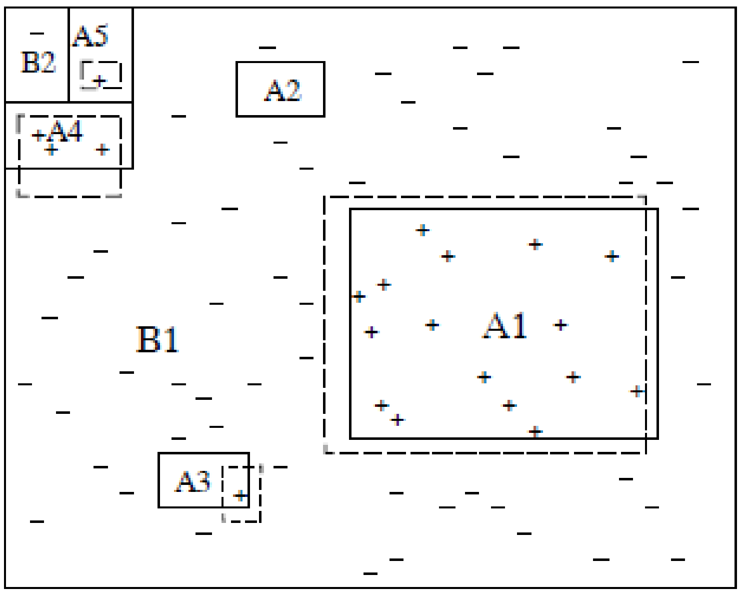


Figure 1: Graphical representation of a rare class and rare case

Class A is the rare (minority class and B is the common (majority class).

Subconcepts A2-A5 correspond to rare cases, whereas A1 corresponds to a fairly common case, covering a substantial portion of the instance space.

Subconcept B2 corresponds to a rare case, demonstrating that common classes may contain rare cases.

G.M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6:1 (2004) 7-19

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Small disjuncts/Rare or excepcional cases

In the real-word domains, rare cases are not easily identified.An approximation is to use a clustering algorithm on each class.

Jo and Japkowicz, 2004: Cluster-based oversampling: A method for inflating small disjuncts.

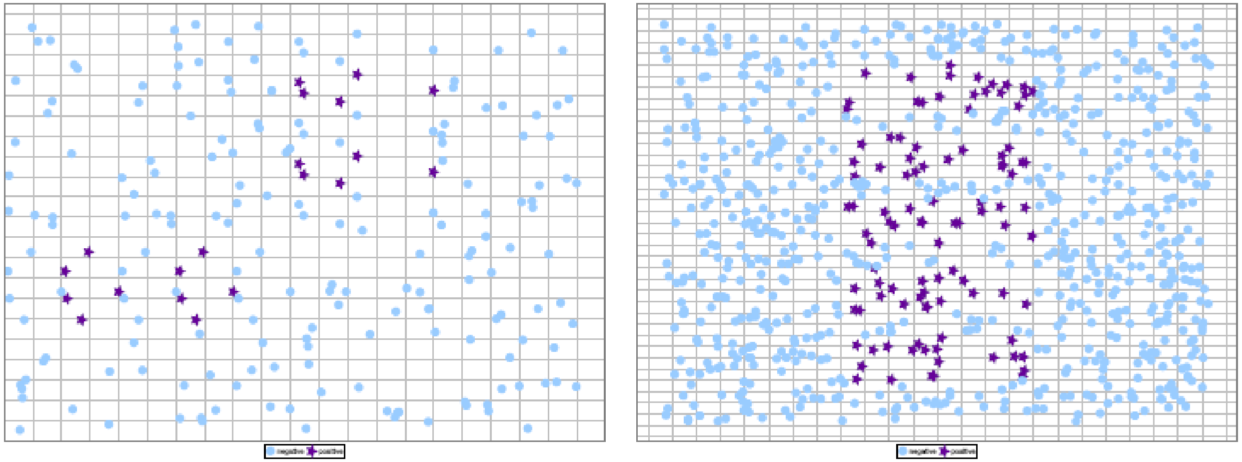
Once the training examples of each class have been clustered, oversampling starts. In the majority class, all the clusters, except for the largest one, are randomly oversampled so as to get the same number of training examples as the largest cluster. Let *maxclasssize* be the overall size of the large class. In the minority class, each cluster is randomly oversampled until each cluster contains *maxclasssize/Nsmallclass* where *Nsmallclass* represents the number of subclusters in the small class.

CBO method: **Cluster-based resampling identifies rare cases and re-samples them individually, so as to avoid the creation of small disjuncts in the learned hypothesis.**

T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Small disjuncts/Rare or excepcional cases



(a) Artificial dataset: small disjuncts for the minority class

(b) Subclus dataset: small disjuncts for both classes

Fig. 5 Example of small disjuncts on imbalanced data

Table 12 Performance obtained by C4.5 in datasets suffering from small disjuncts

Dataset	Original Data			Preprocessed Data with CBO		
	TP_{rate}	TN_{rate}	AUC	TP_{rate}	TN_{rate}	AUC
Artificial dataset	.0000	1.000	.5000	1.000	1.000	1.000
Subclus dataset	1.000	.9029	.9514	1.000	1.000	1.000

T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Rare or excepcional cases

Small disjuncts play a role in the performance loss of class imbalanced domains.
Jo and Japkowicz results show that it is the small disjuncts problem more than the class imbalance problem that is responsible for the this decrease in accuracy.
The performance of classifiers, though hindered by class imbalanced, is repaired as the training set size increases.

An open question: Whether it is more effective to use solutions that address both the class imbalance and the small disjunct problem simultaneously than it is to use solutions that address the class imbalance problem or the small disjunct problem, alone.

T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

Bordeline and Noise data

Dataset shift

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Density: Lack of data

Table 5. The Distribution of Training Examples in Pima Indian Diabetes

		Positive ('1')	Negative ('0')
1:9	40	4	36
	100	10	90
	200	20	180
1:3	40	10	30
	100	25	75
	200	50	150
1:1	40	20	20
	100	50	50
	200	100	100

Different levesl of imbalance and density

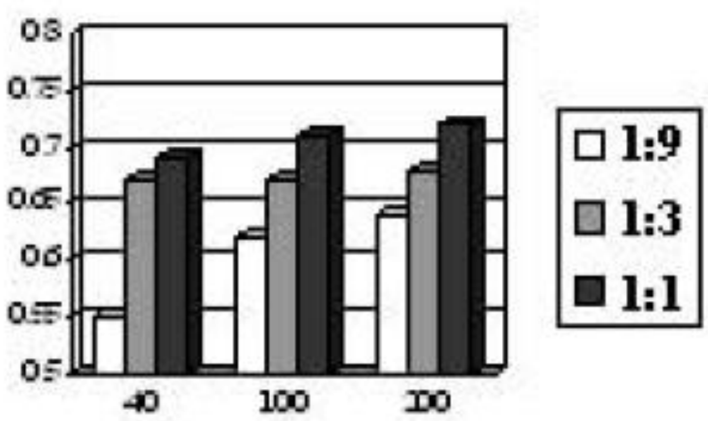
T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

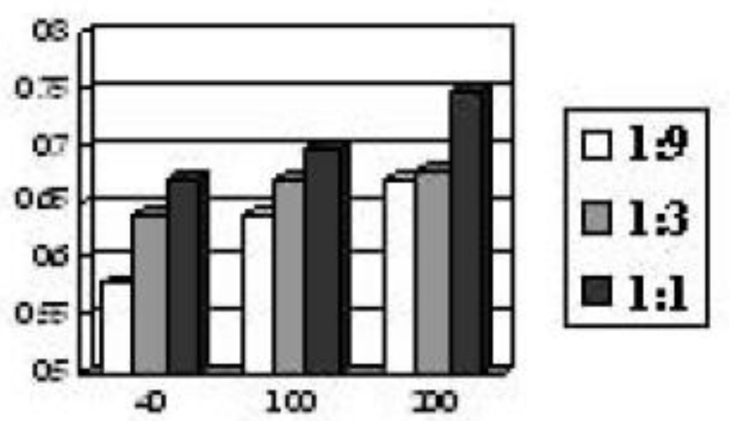
Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Density: Lack of data

Left-C4.5, right-Backpropagation: These resultas show that the performance of classifiers, though hindered by class imbalances, is repaired as the training set size incresses. This sugests that small disjuncts play a role in the performance loss of class imbalanced domains.





T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Density: Lack of data

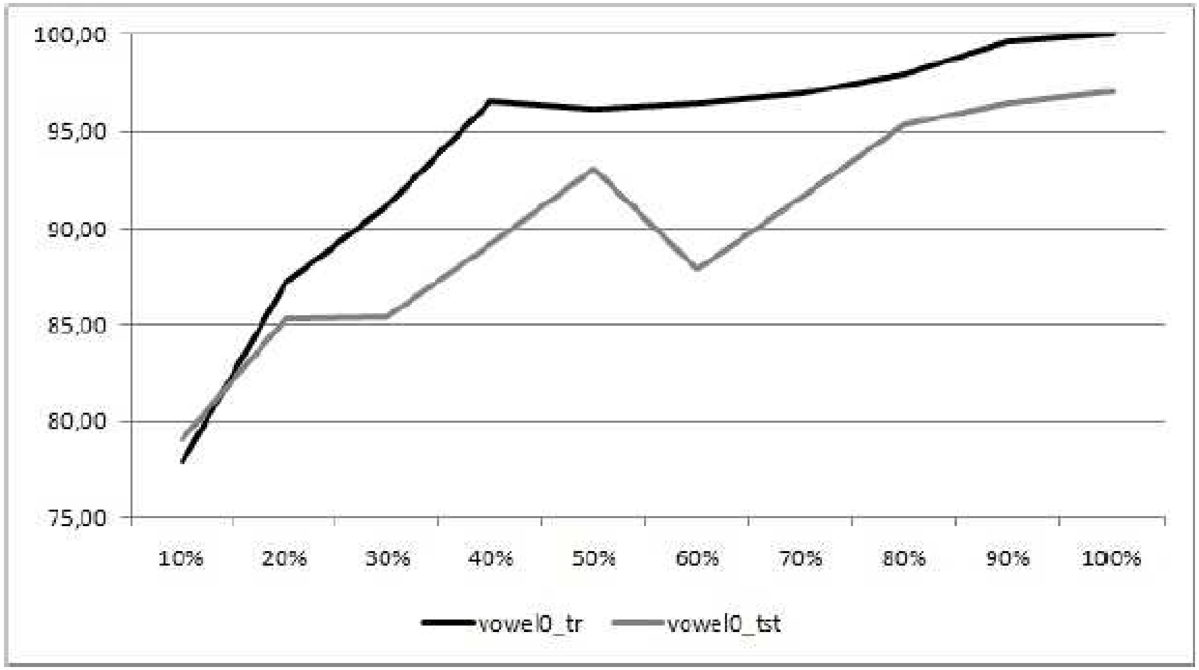


Fig. 8 AUC performance for the C4.5 classifier regarding the proportion of examples in the training set for the vowel0 problem

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

Bordeline and Noise data

Dataset shift

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

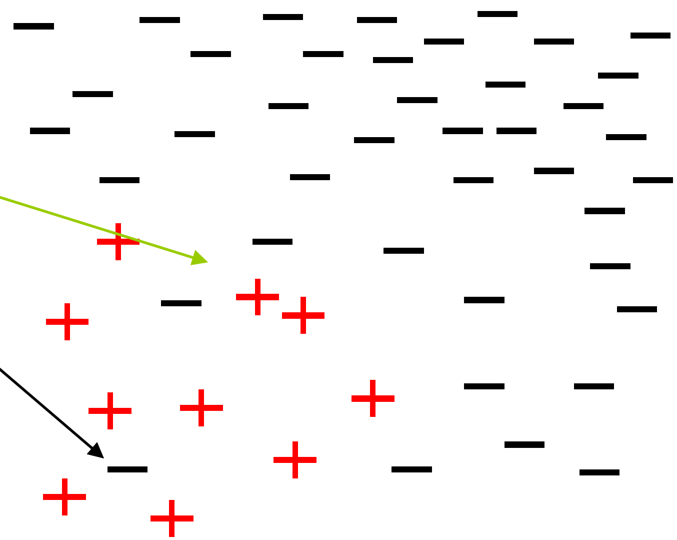
Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Kind of examples: **The need of resampling or to manage the overlapping with other strategies**

- ❑ **Noise examples**
- ❑ **Borderline examples**

  Borderline examples are unsafe since a small amount of noise can make them fall on the wrong side of the decision border.
- ❑ **Redundant examples**

- ❑ **Safe examples**

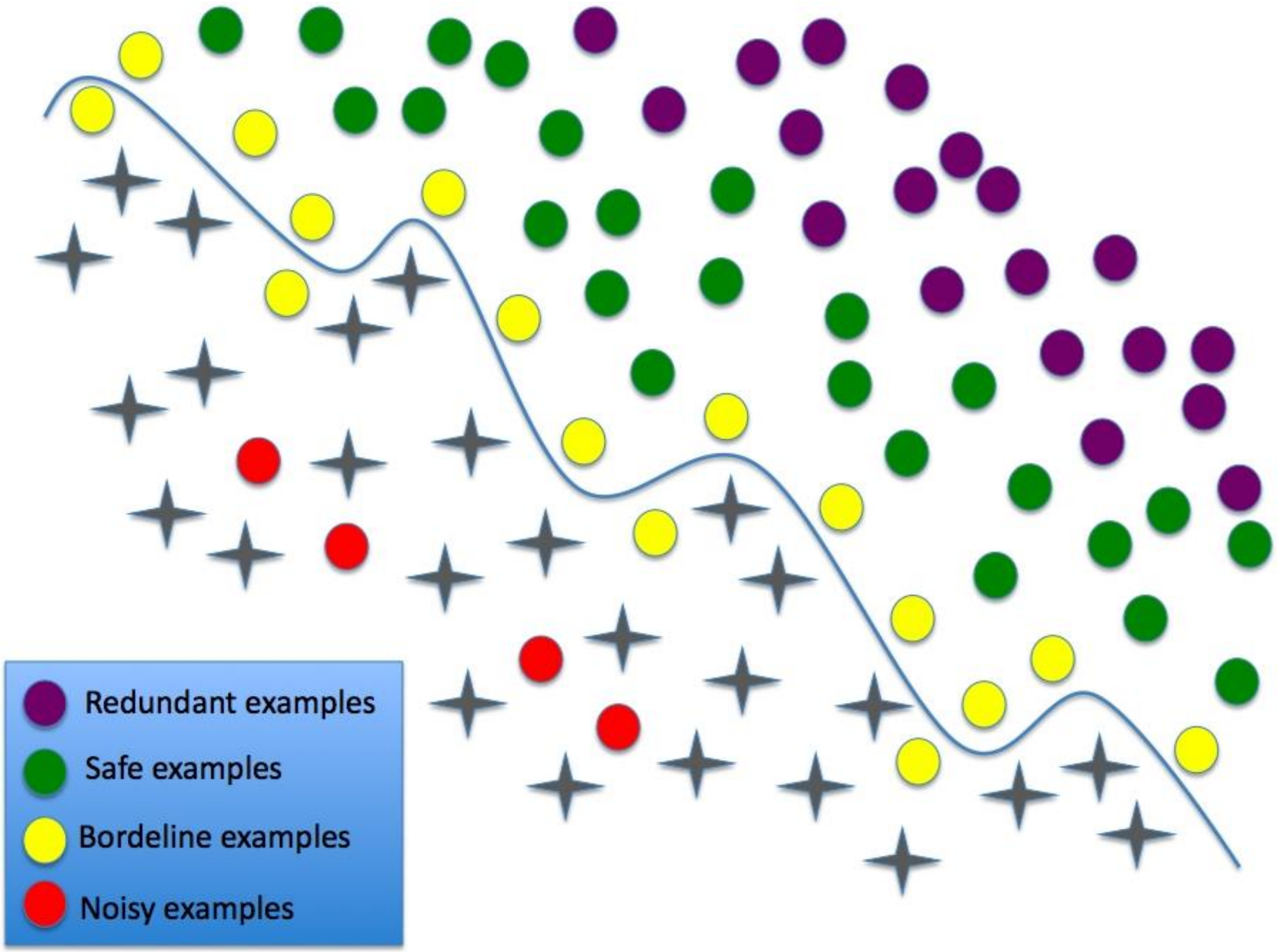


An approach: Detect and remove such majority noisy and borderline examples in filtering before inducing the classifier.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Bordeline and Noise data



Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

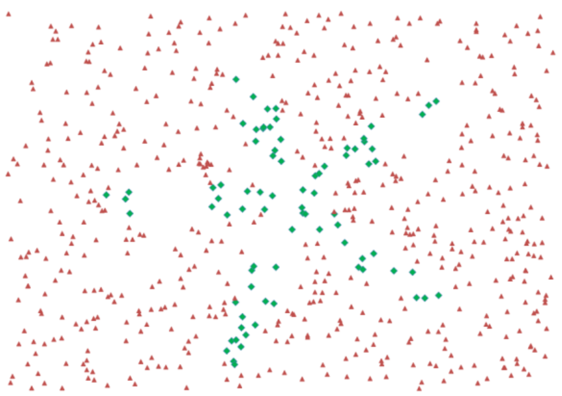
Bordeline and Noise data

3 kind of artificial problems:

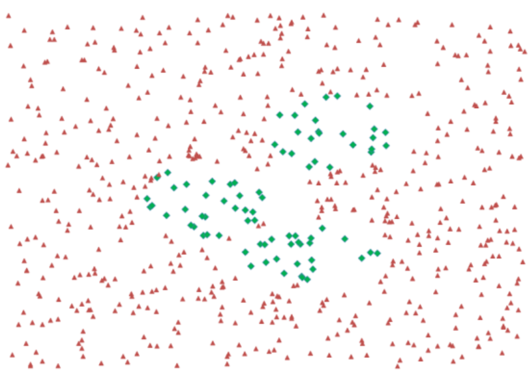
Subclus: examples from the minority class are located inside rectangles following related works on small disjuncts.

Clover: It represents a more difficult, non-linear setting, where the minority class resembles a flower with elliptic petals.

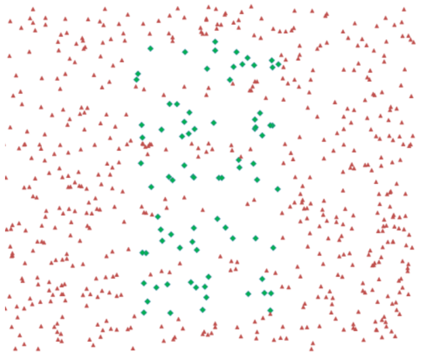
Paw: The minority class is decomposed into 3 elliptic sub-regions of varying cardinalities, where two subregions are located close to each other, and the remaining smaller sub-region is separated.



Clover data



Paw data

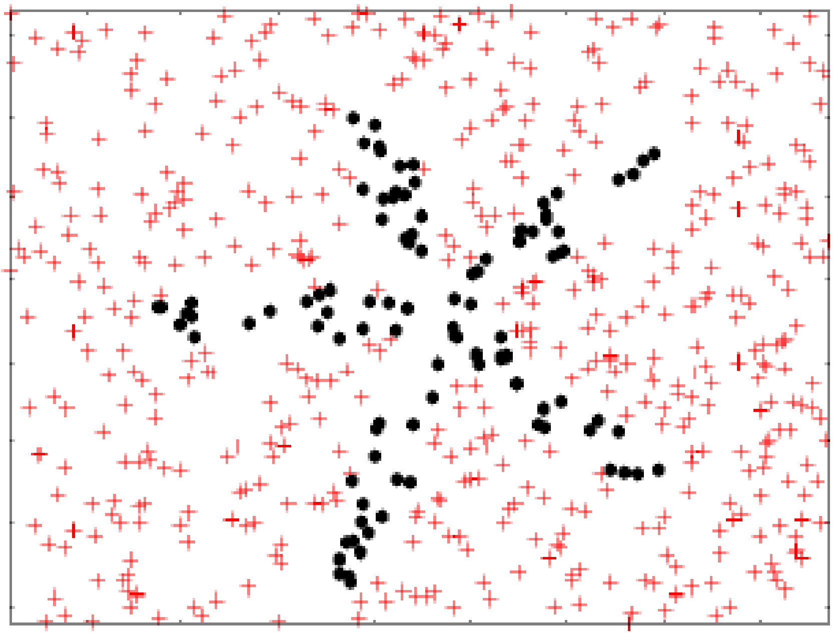


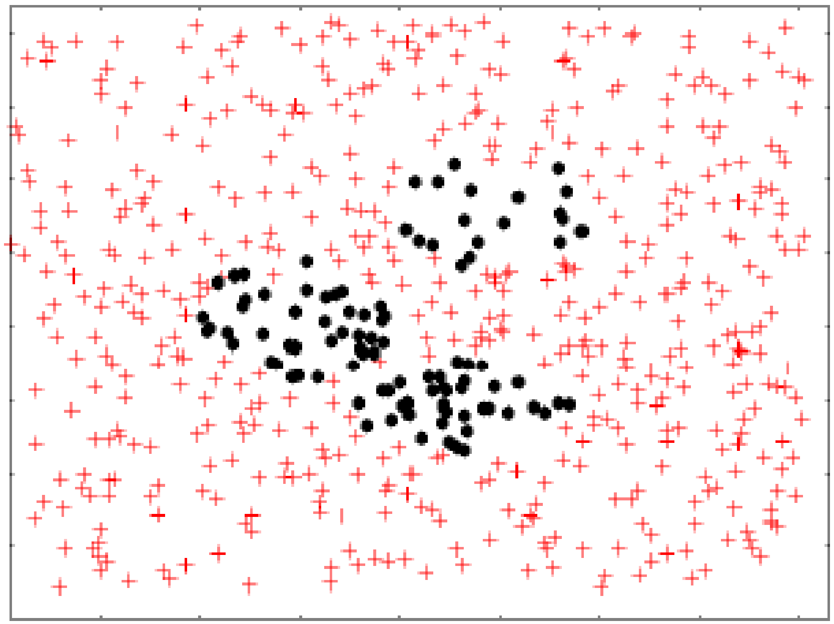
Subclus data

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Bordeline and Noise data





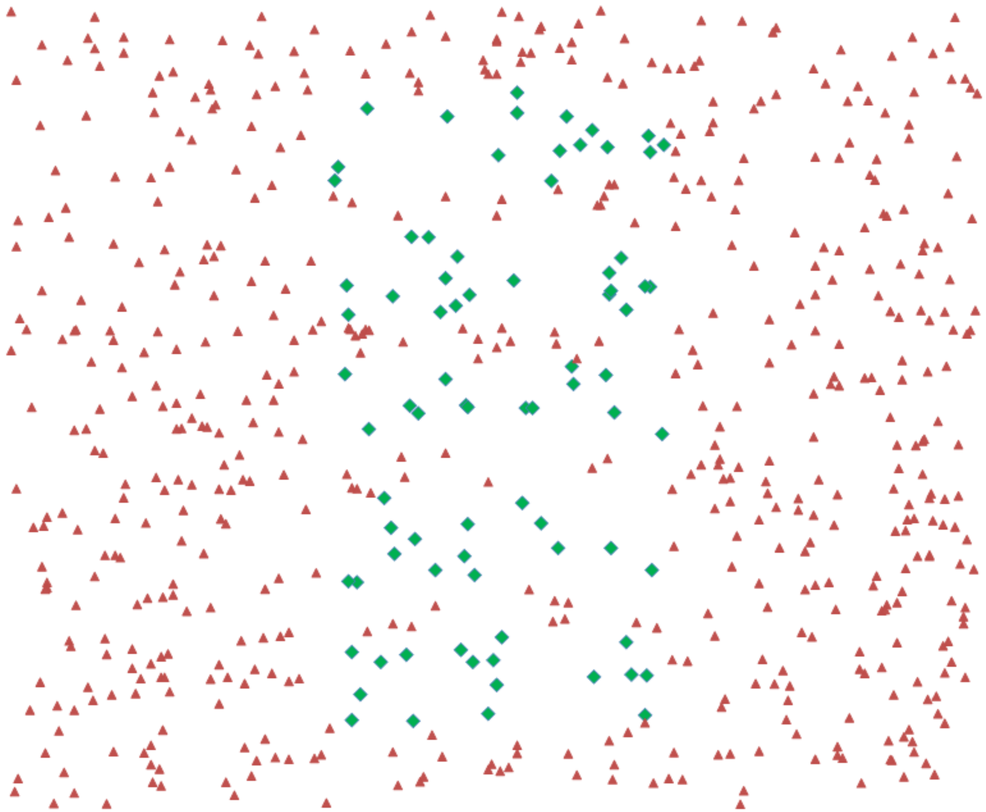
Clover data

Paw data

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

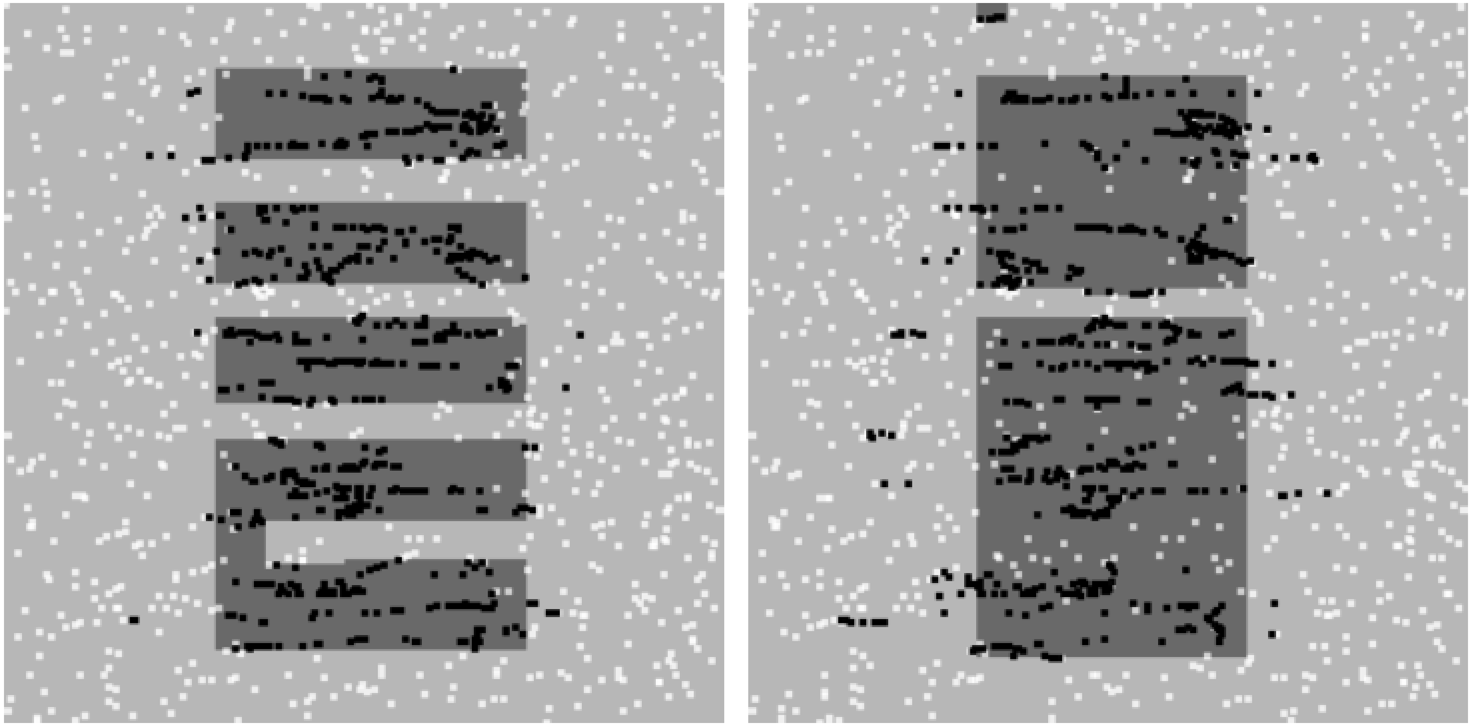
Bordeline and Noise data



Subclus data

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Bordeline and Noise data



(a) Original problem and decision functions (b) Noisy instances and new undesirable decision functions

Fig. 10 Example of the effect of noise in imbalanced datasets for SMOTE+C4.5 in the Subclus dataset

Subclus data

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Bordeline and Noise data

SPIDER 2: Spider family (Selective Preprocessing of Imbalanced Data) rely on the local characteristics of examples discovered by analyzing their k-nearest neighbors.

J. Stefanowski, S. Wilk. Selective pre-processing of imbalanced data for improving classification performance. 10th International Conference in Data Warehousing and Knowledge Discovery (DaWaK2008). LNCS 5182, Springer 2008, Turin (Italy, 2008) 283-292.

K.Napierala, J. Stefanowski, and S. Wilk. **Learning from Imbalanced Data in Presence of Noisy and Borderline Examples**. **7th International Conference on Rough Sets and Current Trends in Computing , 7th International Conference on Rough Sets and Current Trends in Computing,** RSCTC 2010, LNAI 6086, pp. 158–167, 2010.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Bordeline and Noise data

Data set	C4.5 Base	RO	CO	NCR	SP2
subclus-0	0.9540	0.9500	0.9500	0.9460	0.9640
subclus-30	0.4500	0.6840	0.6720	0.7160	0.7720
subclus-50	0.1740	0.6160	0.6000	0.7020	0.7700
subclus-70	0.0000	0.6380	0.7000	0.5700	0.8300
clover-0	0.4280	0.8340	0.8700	0.4300	0.4860
clover-30	0.1260	0.7180	0.7060	0.5820	0.7260
clover-50	0.0540	0.6560	0.6960	0.4460	0.7700
clover-70	0.0080	0.6340	0.6320	0.5460	0.8140
paw-0	0.5200	0.9140	0.9000	0.4900	0.5960
paw-30	0.2640	0.7920	0.7960	0.8540	0.8680
paw-50	0.1840	0.7480	0.7200	0.8040	0.8320
paw-70	0.0060	0.7120	0.6800	0.7460	0.8780

Noise data

Table 14 Performance obtained by C4.5 in the Subclus dataset with and without noisy instances

Dataset	Original Data			20% of Gaussian Noise		
	TP_{rate}	TN_{rate}	AUC	TP_{rate}	TN_{rate}	AUC
None	1.000	.9029	.9514	.0000	1.000	.5000
RandomUnderSampling	1.000	.7800	.8900	.9700	.7400	.8550
SMOTE	.9614	.9529	.9571	.8914	.8800	.8857
SMOTE+ENN	.9676	.9623	.9649	.9625	.9573	.9599
SPIDER2	1.000	1.000	1.000	.9480	.9033	.9256

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Bordeline and Noise data

Small disjunct and Noise data

Bordeline and Noise data

Bordeline and Noise data

Data set	C4.5 Base	RO	CO	NCR	SP2
subclus-0	0.9540	0.9500	0.9500	0.9460	0.9640
subclus-30	0.4500	0.6840	0.6720	0.7160	0.7720
subclus-50	0.1740	0.6160	0.6000	0.7020	0.7700
subclus-70	0.0000	0.6380	0.7000	0.5700	0.8300
clover-0	0.4280	0.8340	0.8700	0.4300	0.4860
clover-30	0.1260	0.7180	0.7060	0.5820	0.7260
clover-50	0.0540	0.6560	0.6960	0.4460	0.7700
clover-70	0.0080	0.6340	0.6320	0.5460	0.8140
paw-0	0.5200	0.9140	0.9000	0.4900	0.5960
paw-30	0.2640	0.7920	0.7960	0.8540	0.8680
paw-50	0.1840	0.7480	0.7200	0.8040	0.8320
paw-70	0.0060	0.7120	0.6800	0.7460	0.8780

Table 14 Performance obtained by C4.5 in the Subclus dataset with and without noisy instances

Dataset	Original Data			20% of Gaussian Noise		
	TP_{rate}	TN_{rate}	AUC	TP_{rate}	TN_{rate}	AUC
None	1.000	.9029	.9514	.0000	1.000	.5000
RandomUnderSampling	1.000	.7800	.8900	.9700	.7400	.8550
SMOTE	.9614	.9529	.9571	.8914	.8800	.8857
SMOTE+ENN	.9676	.9623	.9649	.9625	.9573	.9599
SPIDER2	1.000	1.000	1.000	.9480	.9033	.9256

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Bordeline and Noise data

- SPIDER 2: allows to get good results in comparison with classical ones.

- It has interest to analyze the use of noise filtering algorithms for these problems: IPF filtering algorithm shows good results.

José A. Sáez, J. Luengo, Jerzy Stefanowski, F. Herrera, **SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering**. *Information Sciences 291 (2015) 184-203, doi: 10.1016/j.ins.2014.08.051*.

- Specific methods for managing the noise and bordeline problems are necessary.

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

Bordeline and Noise data

Dataset shift

Three connected problems

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Small disjuncts and density

Rare cases may be due to a lack of data. Relative lack of data, relative rarity.

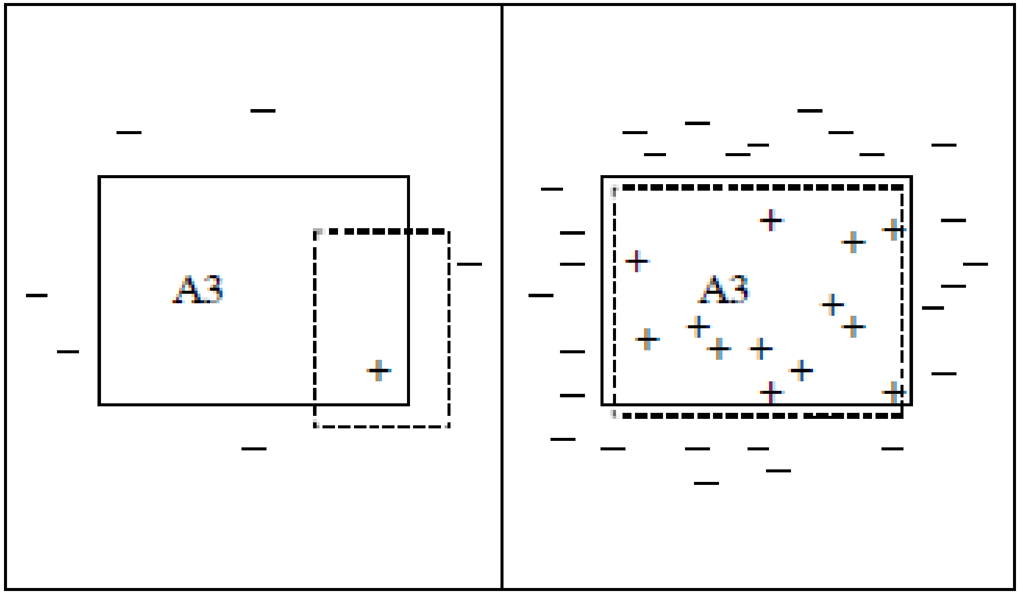


Figure 2: The impact of an "absolute" lack of data

G.M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6:1 (2004) 7-19

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Small disjuncts and Noise data

Noise data will affect the way any data mining system behaves. Noise has a greater impact on rare cases than on common cases.

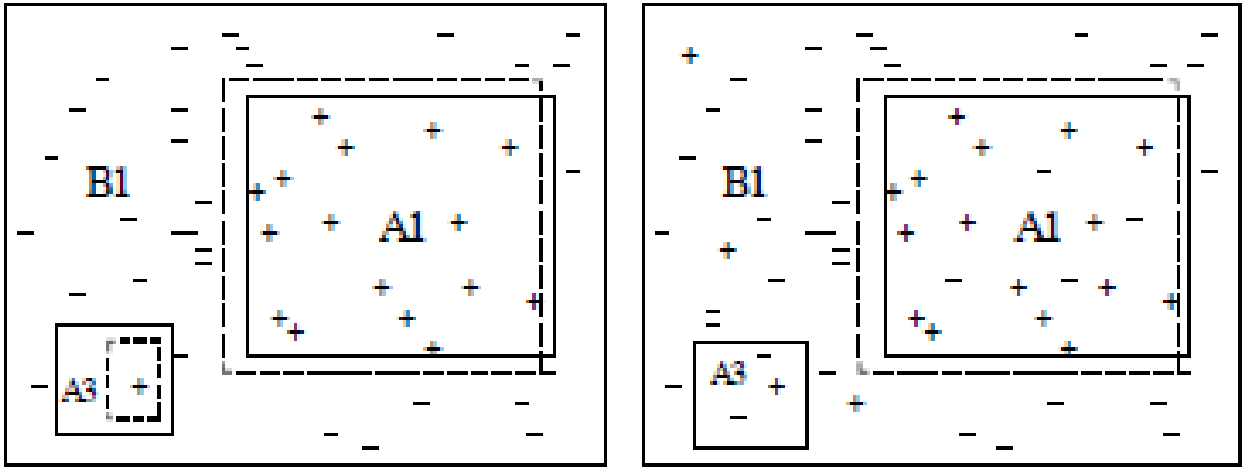


Figure 3: The effect of noise on rare cases

G.M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6:1 (2004) 7-19

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

Bordeline and Noise data

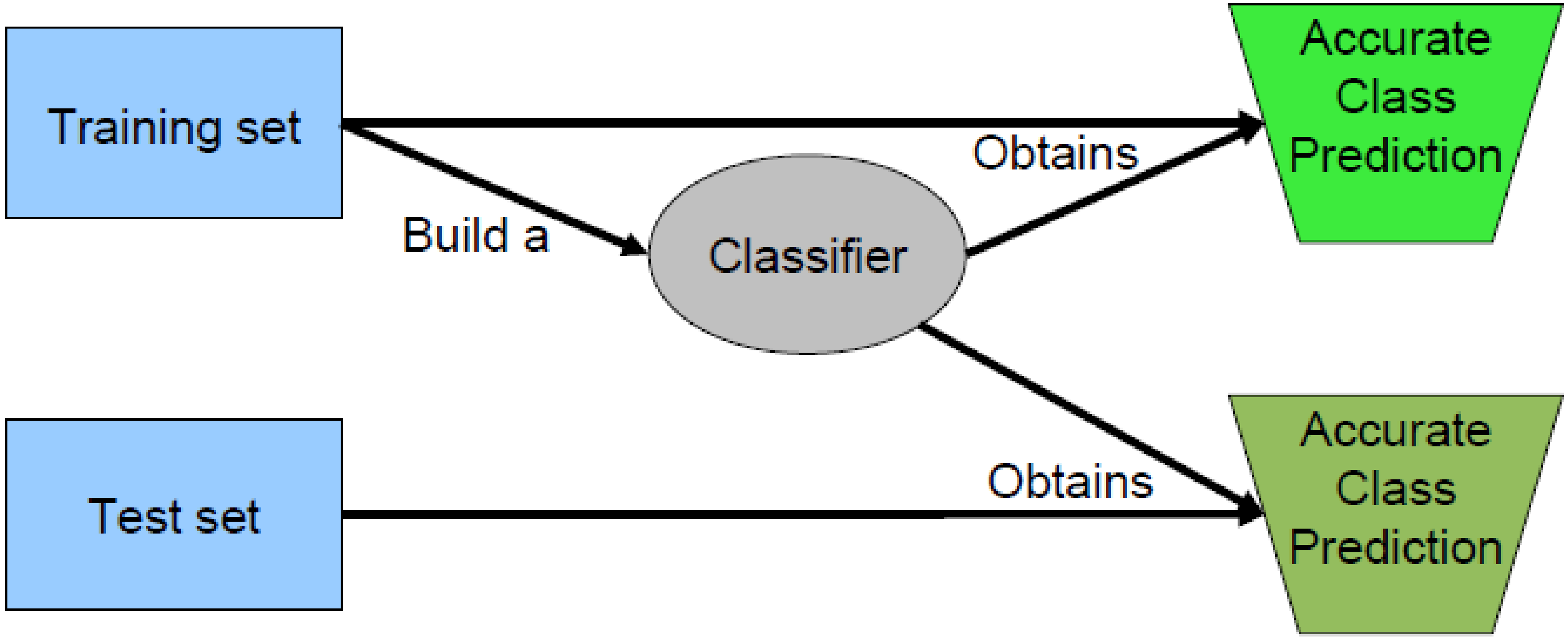
Dataset shift

V. López, A. Fernandez, S. García, V. Palade, F. Herrera, **An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.** Information Sciences 250 (2013) 113-141.

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Dataset shift

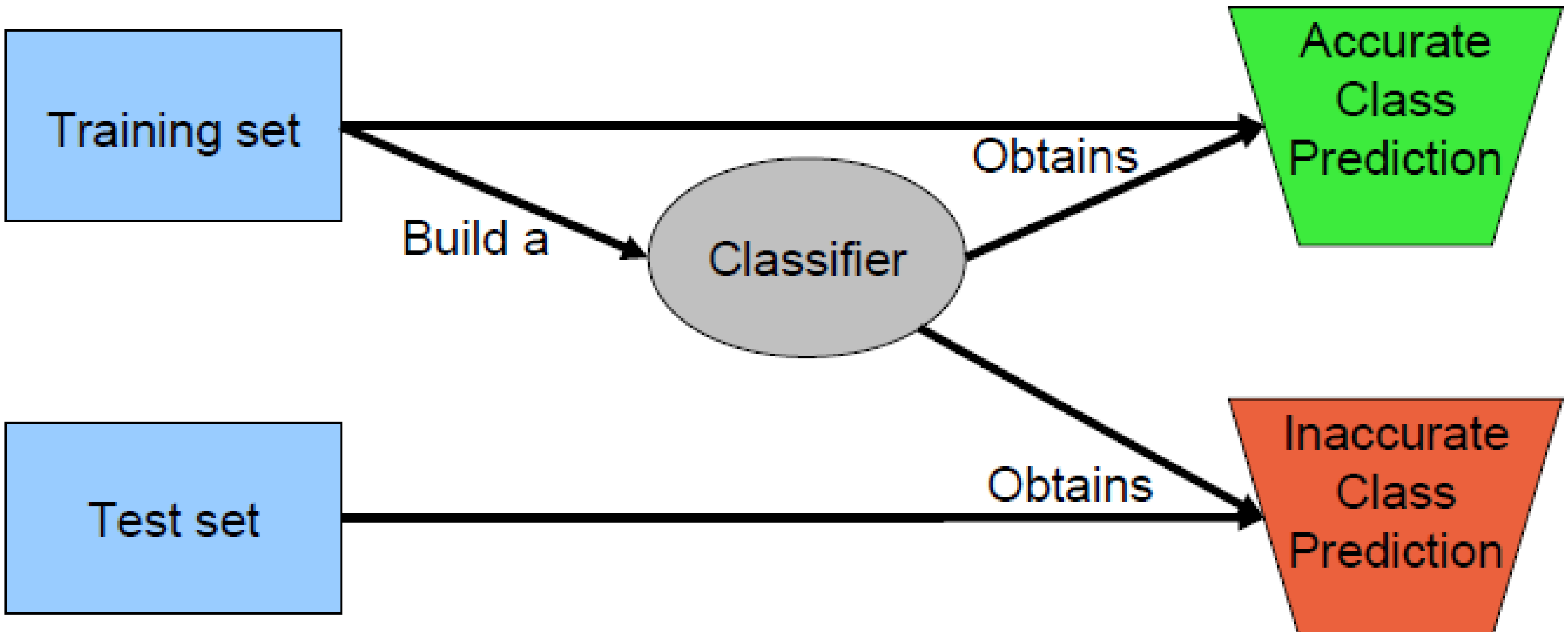
- Basic assumption in classification:



Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Dataset shift

- But sometimes….



- **The classifier has an overfitting problem.**
- **Is there a change in data distribution between training and test sets (Data fracture)?.**

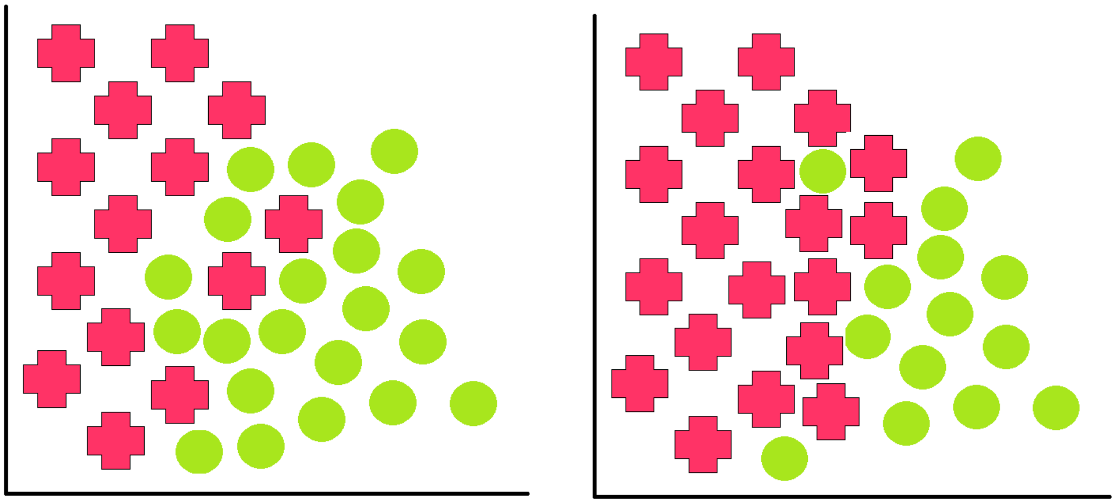
The Problem of Dataset Shift

- **The classifier has an overfitting problem.**
  - **Change the parameters of the algorithm.**
  - **Use a more general learning method.**

- **There is a change in data distribution between training and test sets (Dataset shift).**
  - **Train a new classifier for the test set.**
  - **Adapt the classifier.**
  - **Modify the data in the test set …**

The Problem of Dataset Shift

The problem of data-set shift is defined as the case where training and test data follow different distributions.



J. G. Moreno-Torres, T. R. Raeder, R. Aláiz-Rodríguez, N. V. Chawla, F. Herrera, A unifying view on dataset shift in classification. ***Pattern Recognition 45:1 (2012) 521-530, doi:10.1016/j.patcog.2011.06.019*****.**

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Dataset shift

This is a common problem that can affect all kind of classification problems, and it often appears due to sample selection bias issues.

However, **the data-set shift issue is specially relevant when dealing with imbalanced classification**, because in highly imbalanced domains, the minority class is particularly sensitive to singular classification errors, due to the typically low number of examples it presents.

In the most extreme cases, a single misclassified example of the minority class can create a significant drop in performance.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Dataset shift

Since dataset shift is a **highly relevant issue in imbalanced classification,** it is easy to see why it would be an interesting perspective to focus on future research regarding the topic.

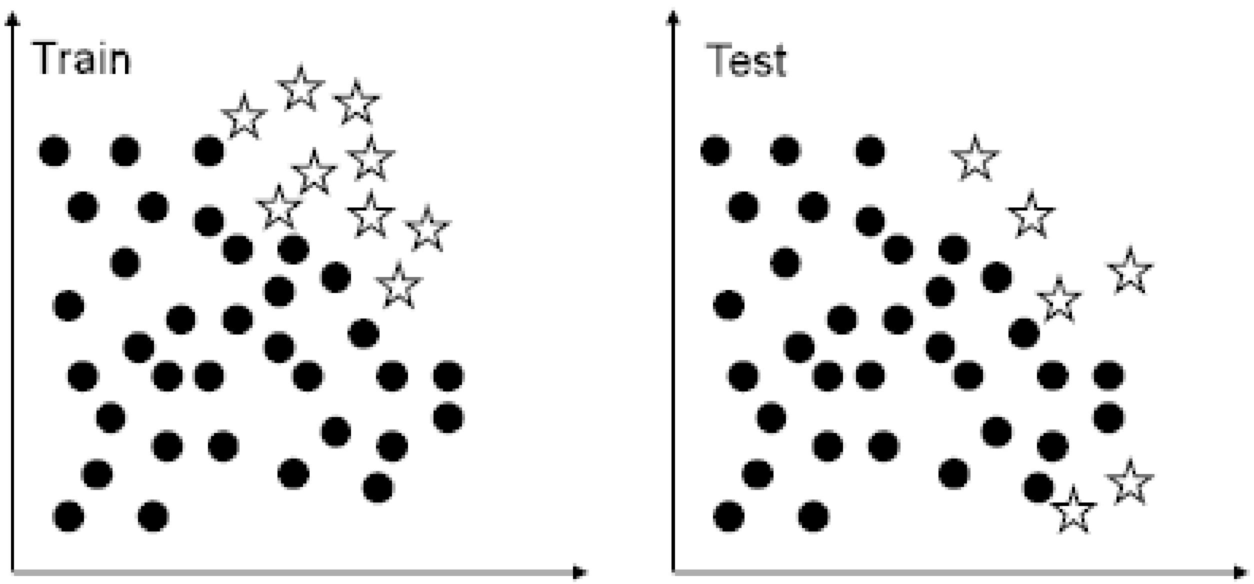


Figure 18: Example of the impact of data-set shift in imbalanced domains.

Causes of Dataset Shift

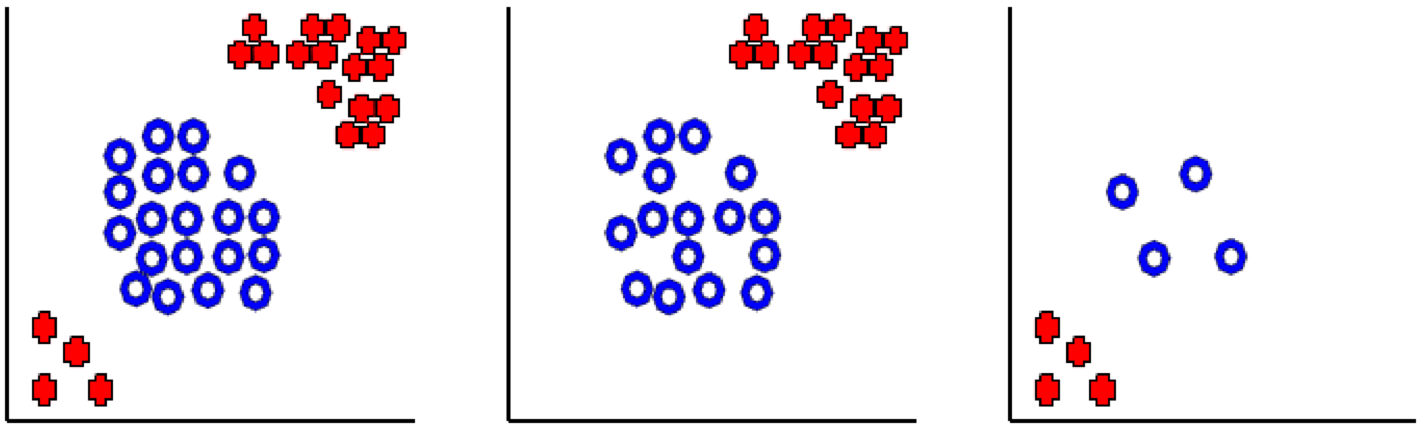
We comment on some of the most common causes of Dataset Shift:

Sample selection bias and non-stationary environments.

These concepts have created confusion at times, so it is important to remark **that these terms are factors that can lead to the appearance of some of the shifts explained, but they do not constitute Dataset Shift themselves.**

Causes of Dataset Shift

Sample selection bias:

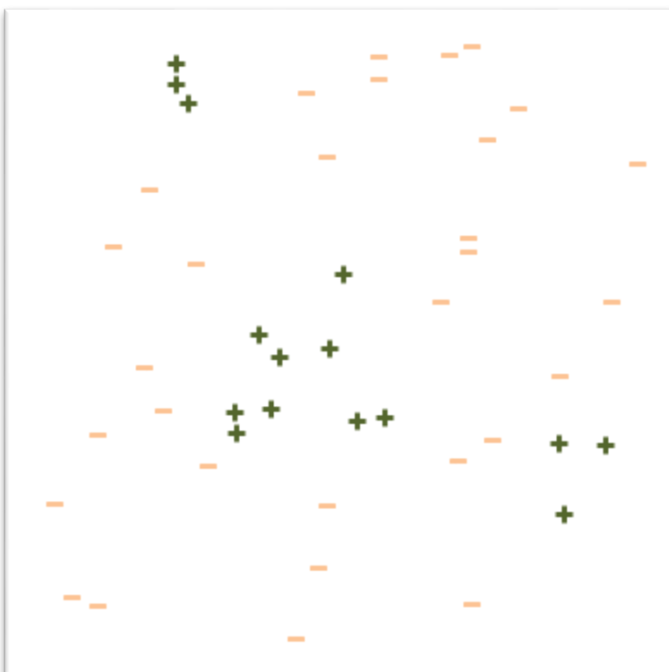


(a) Full dataset (b) Training set (c) Test set

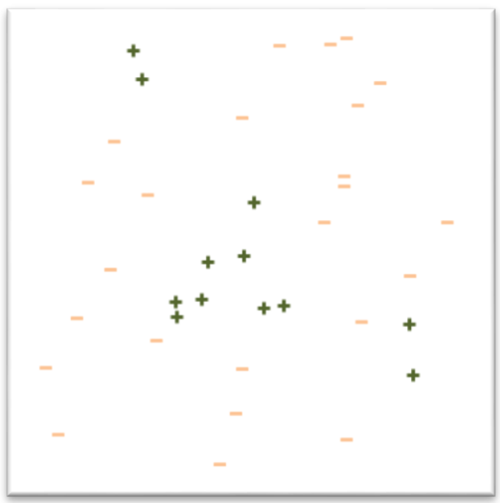
Fig. 1: Extreme example of partition-based covariate shift. Note how the examples on the bottom left of the "cross" class will be wrongly classified due to covariate shift.

Causes of Dataset Shift

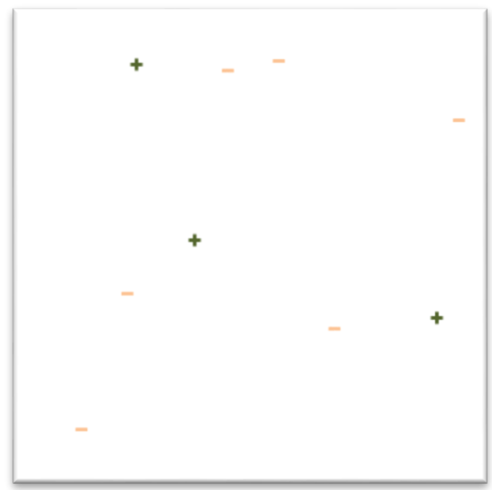
- **Training and test following the same data distribution**



Original Data



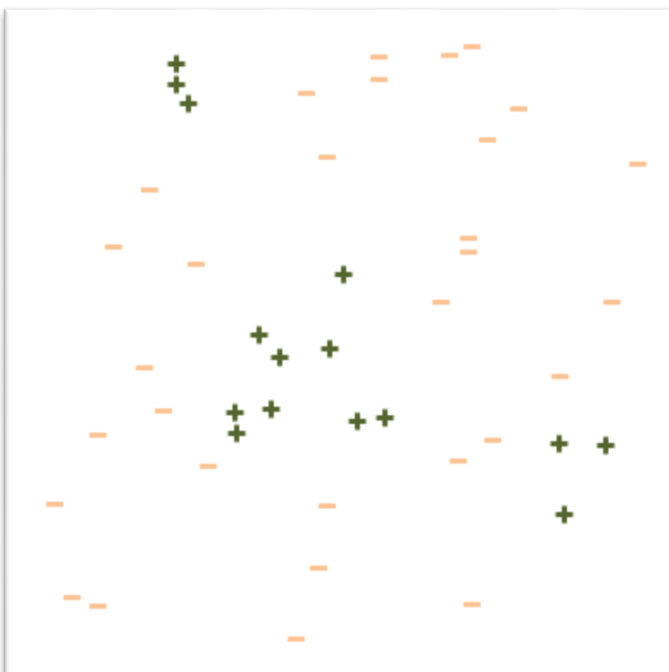
Training Data



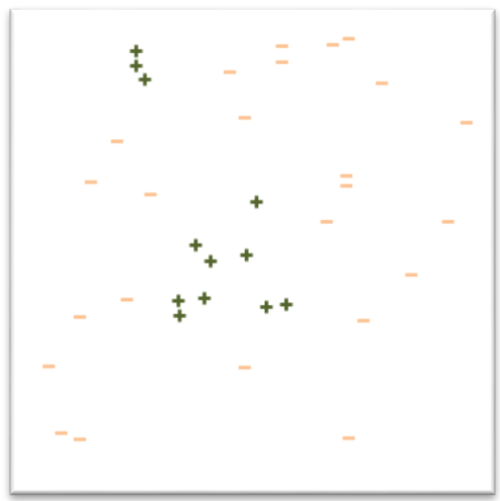
Test Data

Causes of Dataset Shift

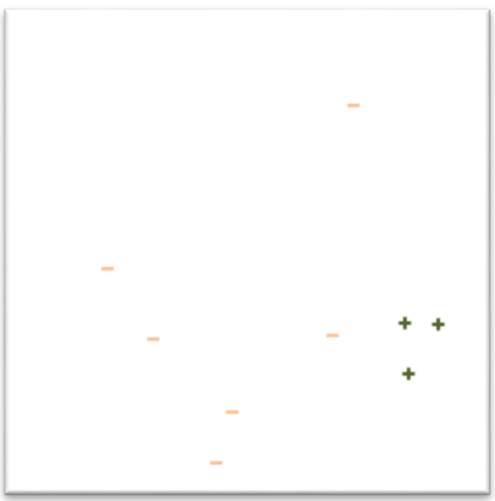
- **DATASET SHIT: Training and test following <u>different</u> data distribution**



Original Data



Training Data



Test Data

Causes of Dataset Shift

Sample bias selection: Influence of partitioning on classifiers' performance

	Iteration 216		Iteration 459	
	C45	HDDT	C45	HDDT
breast-w	**0.9784**	0.9753	0.9768	**0.9820**
bupa	**0.6936**	0.6913	0.6521	**0.6531**
credit-a	**0.8996**	0.8967	**0.9044**	0.8967
crx	**0.8993**	0.8877	**0.9021**	0.8898
heart-c	**0.8431**	0.8181	0.8161	**0.8333**
heart-h	**0.8756**	0.8290	0.8376	**0.8404**
horse-colic	0.8646	**0.8848**	0.8742	**0.8928**
ion	**0.9353**	0.9301	0.9247	**0.9371**
krkp	0.9992	**0.9993**	0.9988	**0.9991**
pima	**0.7781**	0.7717	0.7661	**0.7696**
promoters	**0.8654**	0.8514	0.8676	**0.8774**
ringnorm	**0.8699**	0.8533	0.8669	**0.8727**
sonar	**0.8053**	0.7929	0.8076	**0.8127**
threenorm	**0.7964**	0.7575	**0.7419**	0.7311
tic-tac-toe	**0.9354**	0.9254	**0.9342**	0.9273
twonorm	**0.8051**	0.8023	0.7722	**0.7962**
vote	**0.9843**	0.9824	0.9828	**0.9835**
vote1	**0.9451**	0.9343	**0.9497**	0.9426
avg. rank	**1.11**	1.89	1.72	**1.28**
$\alpha = 0.10$	✓			✓
$\alpha = 0.05$	✓			✓

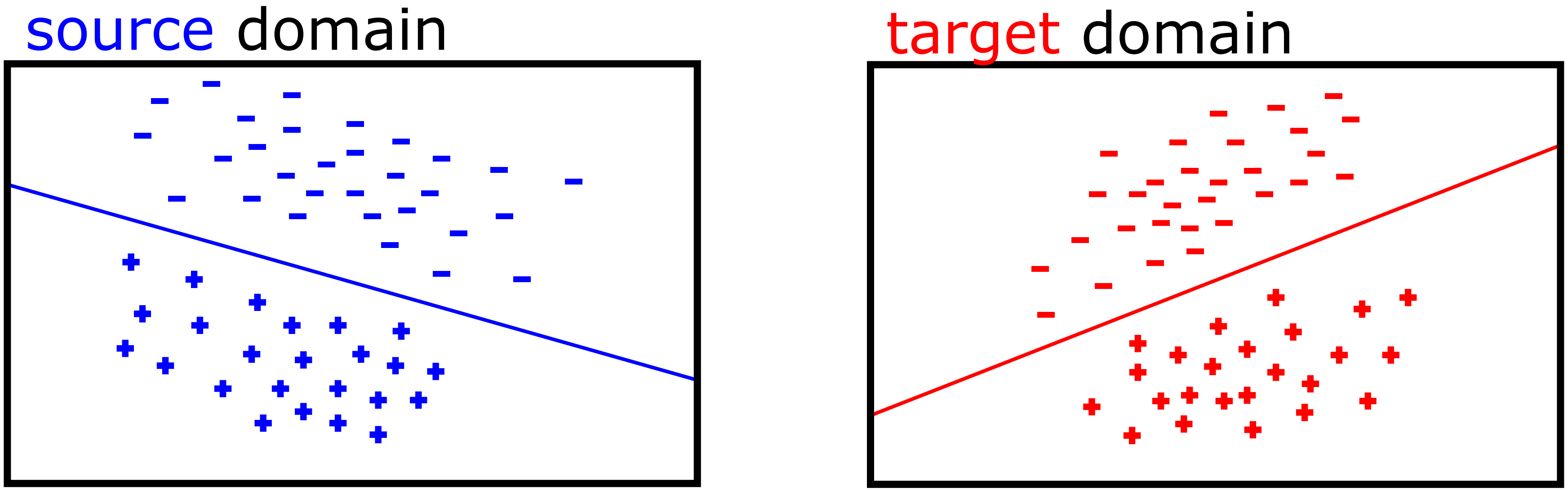
- **Classifier performance results over two separate iterations of random 10-fold cross-validation.**
- **A consistent random number seed was used across al datasets within an iteration.**

T. Raeder, T. R. Hoens, and N. V. Chawla, "Consequences of variability in classifier performance estimates," Proceedings of the 2010 IEEE International Conference on Data Mining, 2010, pp. 421–430.

Wilcoxon test: Clear differences for both algorithms

Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

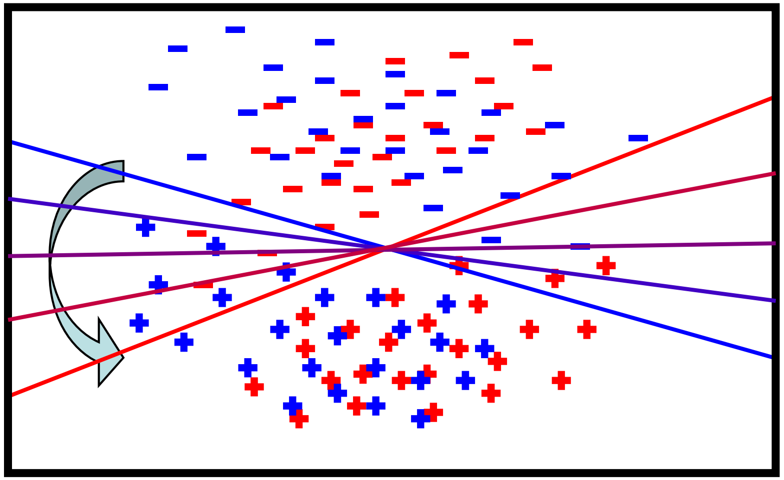


Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

source domain

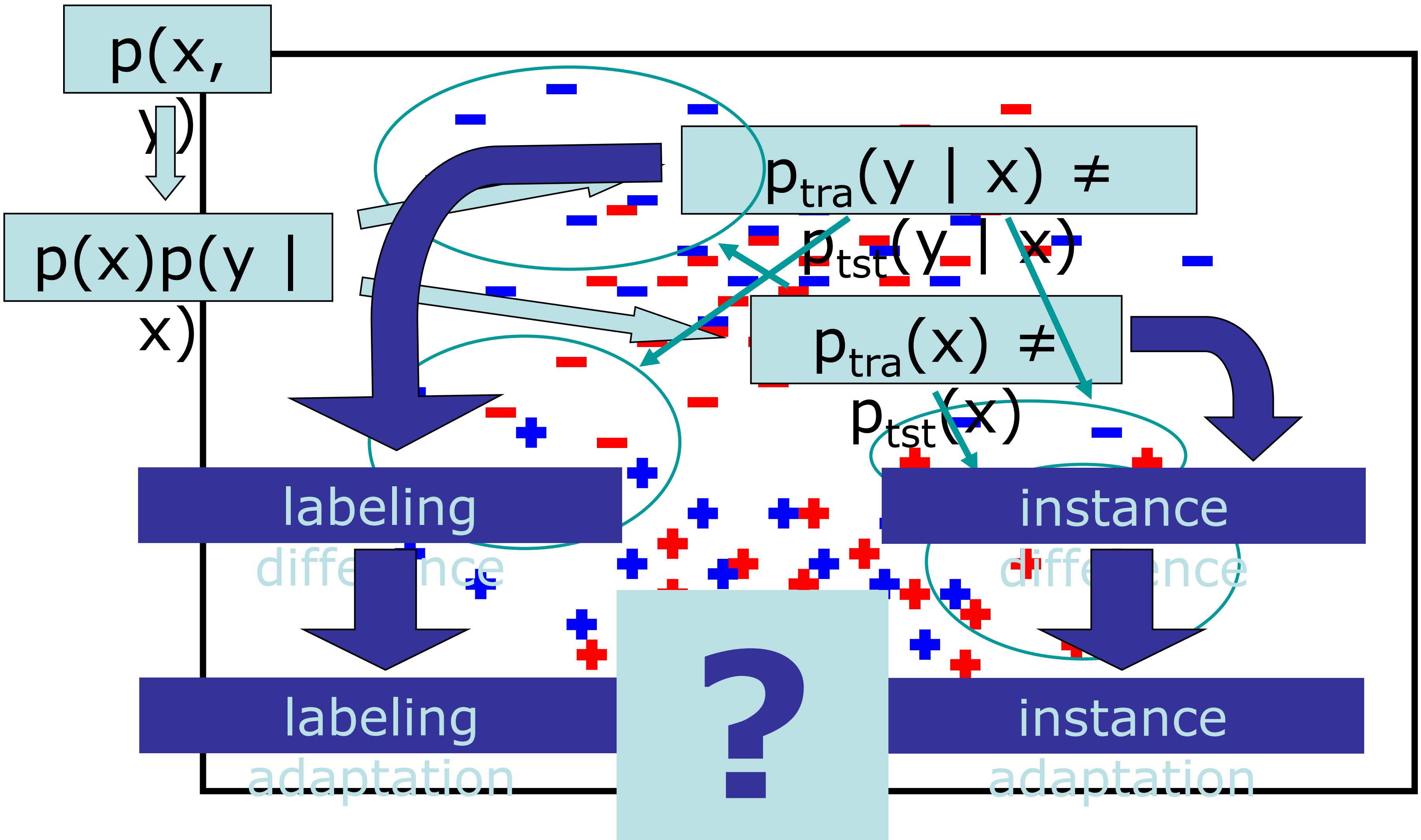
target domain



Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Where Does the Difference Come from?

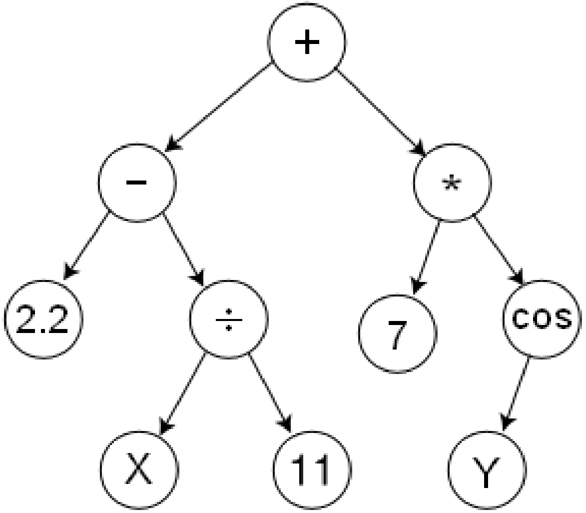


Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Dataset shift

GP-RST: From N dimensions to 2

- Goal: obtain a 2-dimensional representation of a given dataset that is as separable as possible.
- Genetic Programming based: evolves 2 trees simultaneously as arithmetic functions of the previous N-dimensions.
- Evaluation of an individual dependant on Rough Set Theory measures.

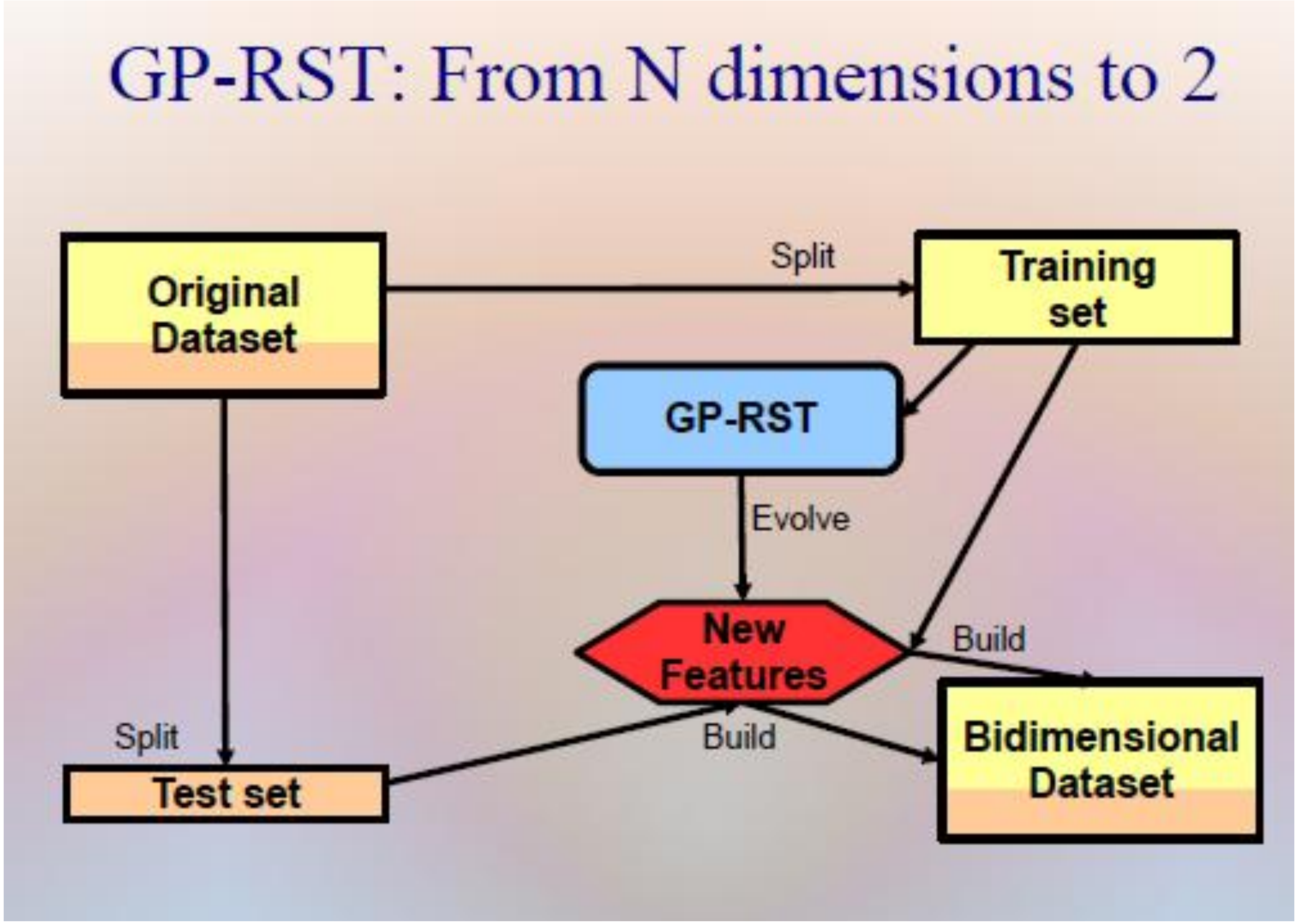


$$\left(2.2 - \left(\frac{X}{11}\right)\right) + \left(7 * \cos(Y)\right)$$

Moreno-Torres, J. G., & Herrera, F. (2010). A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA 2010) (pp. 501–506).*

Why is difficult to learn in imbalanced domains? Intrinsic data characteristics

Data-set shift



A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

The quality of approxim $\gamma(x)$ n is the proportion of the elements of a rough set that belong to its lower approximation.

$$B_*(X) = \{x \in X : R'(x) \subseteq X\}$$

$$\gamma(x) = \frac{|B_*(X)|}{|X|}$$

Algorithm 1 Fitness evaluation procedure

1. Obtain $E' = \{e'^h = (f_1(e^h), f_2(e^h), C^h)/h = 1, ..., n_e\}$, where f_1 and f_2 are the expressions encoded on each of the trees of the individual being evaluated.
2. For each class label $C_i \in C : i = 1, ..., n_c$,
   - 2.1 Build a rough set X_i containing all the elements of class C_i.
   - 2.2 Calculate the lower approximation of X_i, $B_*(X_i)$.
   - 2.3 The fitness of the chromosome for class C_i is estimated as the quality of the approximation over X_i, $\gamma(X_i)$.
3. The fitness of the chromosome is the geometric mean of the ones obtained for each class: $fitness = \sqrt[n_c]{\prod_{i=1}^{n_c} \gamma(X_i)}$.

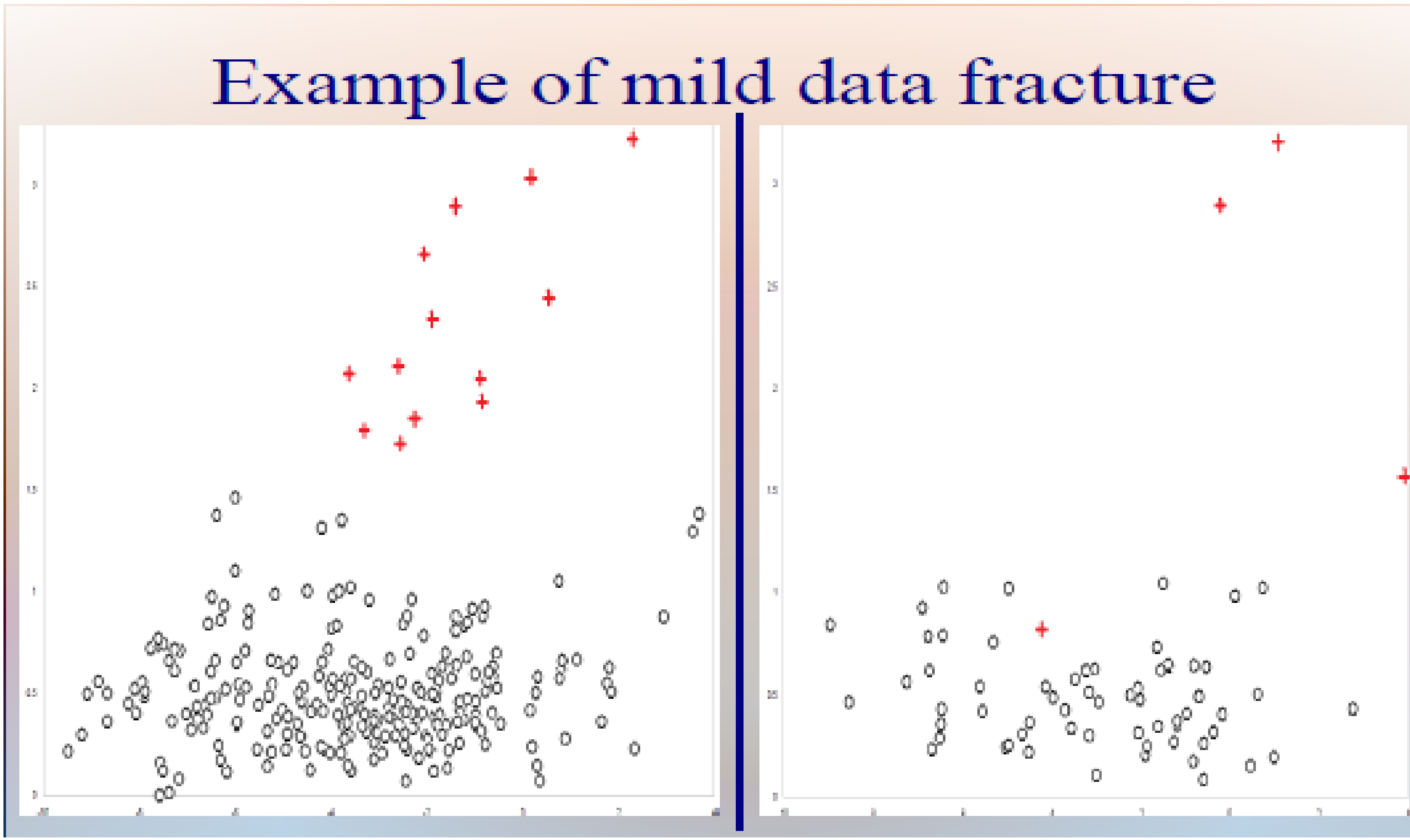
A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Good behaviour. pageblocks 13v4, 1st partition.

(a) Training set (1.0000) (b) Test set (1.0000)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

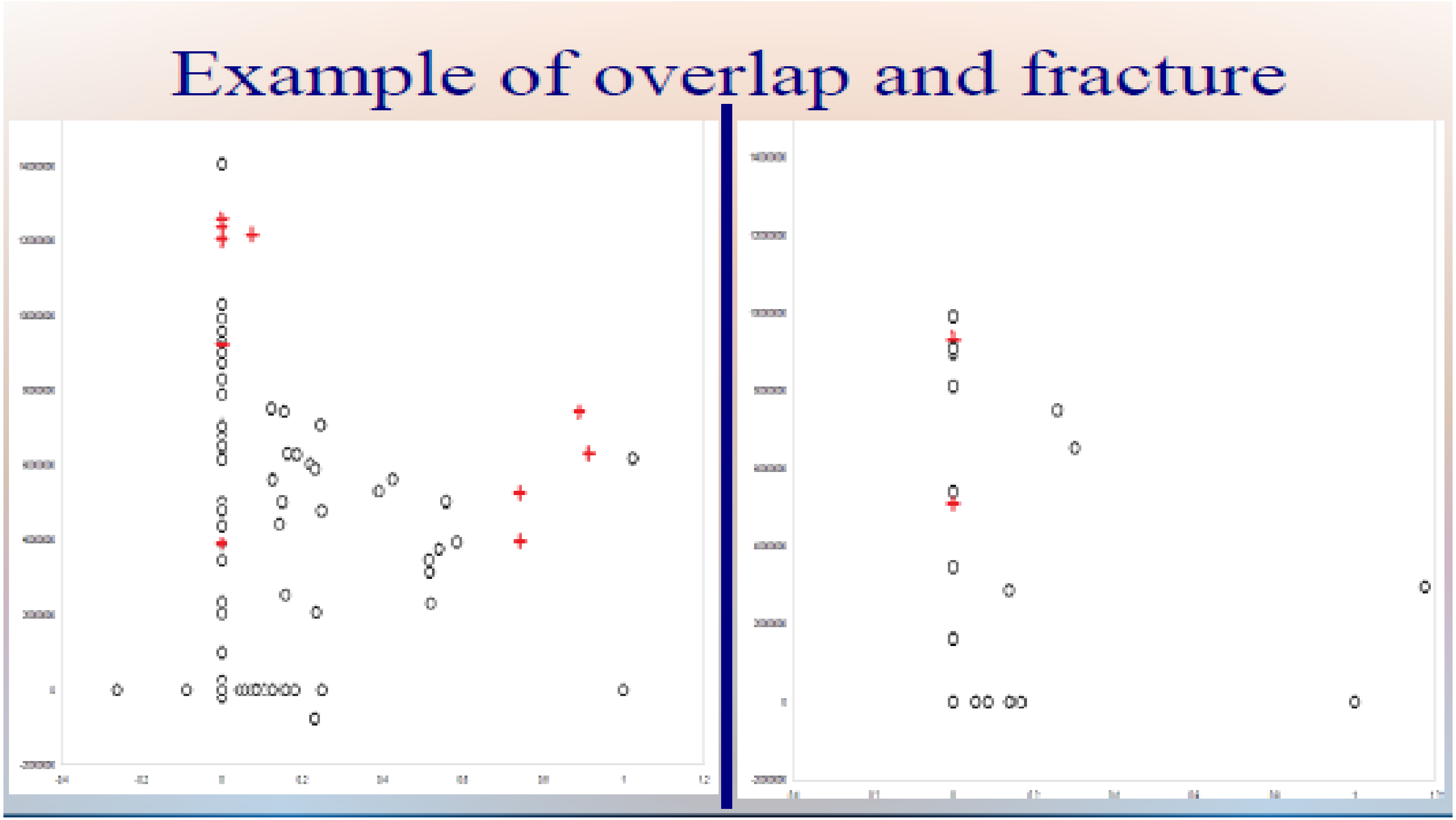
Dataset shift. ecoli 4, 1st partition.



(a) Training set (0.9663)   (b) Test set (0.8660)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

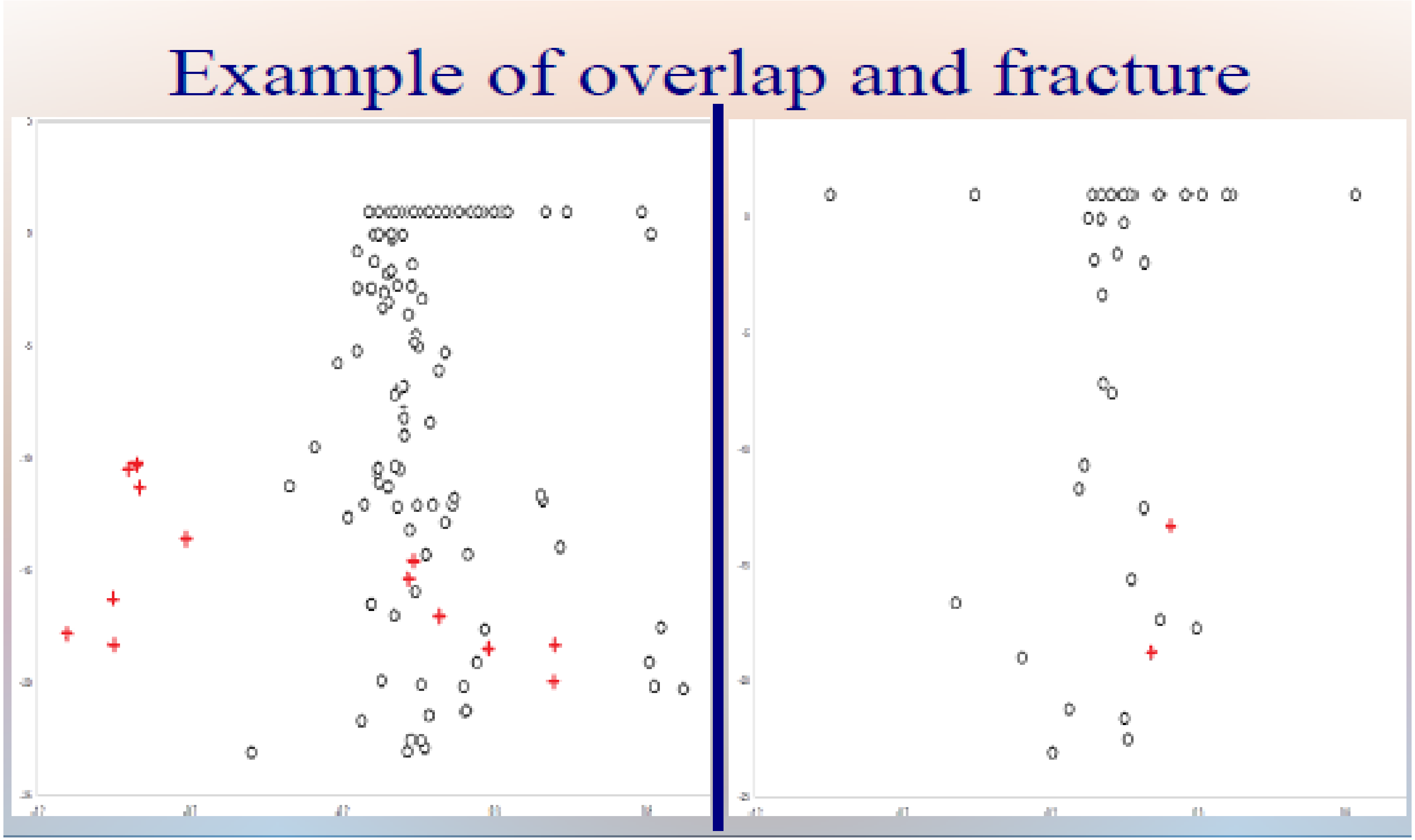
Overlap and dataset shift. glass 016v2, 4th partition.



(a) Training set (0.3779) (b) Test set (0.0000)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Overlap and dataset shift. glass 2, 2nd partition



(a) Training set (0.6794) (b) Test set (0.0000)

Why is difficult to learn in imbalanced domains?
Intrinsic data characteristics

Dataset shift

There are two different potential approaches in the study of the effect and solution of data-set shift in imbalanced domains.

- The first one focuses on intrinsic data-set shift, that is, the data of interest includes some degree of shift that is producing a relevant drop in performance. In this case, we need to:
  - Develop techniques to discover and measure the presence of data-set shift adapting them to minority classes.
  - Design algorithms that are capable of working under data-set shift conditions. These could be either preprocessing techniques or algorithms that are designed to have the capability to adapt and deal with dataset shift without the need for a preprocessing step.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Data-set shift

- The second branch in terms of data-set shift in imbalanced classification is related to induced data-set shift.
Most current state of the art research is validated through stratified cross-validation techniques, which are another potential source of shift in the machine learning process.

A more suitable validation technique needs to be developed in order to avoid introducing data-set shift issues artificially.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

Dataset shift

- Imbalanced classification problems are difficult when overlap and/or data fracture are present.
- Single outliers can have a great influence on classifier performance.
- This is a novel problem in imbalanced classification that need a lot of studies.

Why is difficult to learn in imbalanced domains?

Intrinsic data characteristics

What domain characteristics aggravate the problem?

- ❑Overlapping
- ❑ Rare sets/ Small disjuncts: **The class imbalance problem may not be a problem in itself. Rather, the small disjunct problem it causes is responsible for the decay.**
- ❑ The overall size of the training set
  - large training sets yield low sensitivity to class imbalances
- ❑ Noise and border data provokes additional problems.
- ❑ An increase in the degree of class imbalance. The data partition provokes data fracture: Dataset shift.

Contents

Class Imbalance: Data sets, implementations, …

KEEL Data Mining Tool:
It includes algorithms and data set partitions

http://www.keel.es

KEEL-dataset
Data set repository

KEEL
KNOWLEDGE EXTRACTION based on EVOLUTIONARY LEARNING

Class Imbalance: Data sets, implementations, …

- KEEL is an open source (GPLv3) Java software tool to assess evolutionary algorithms for Data Mining problems including regression, classification, clustering, pattern mining and so on.

- It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques.
- It includes a large list of algorithms for imbalanced data.

Imbalanced Classification (42)	Resampling Data Space (20)	Over-sampling Methods (12)
		Under-sampling Methods (8)
	Cost-Sensitive Classification (3)	
	Ensembles for Class Imbalance (19)	

Class Imbalance: Data sets, implementations, ...

- We include 111 data sets:
  66 for 2 classes,
  15 for multiple classes and
  30 for noise and bordeline.

KEEL-dataset
Data set repository

Imbalanced data sets

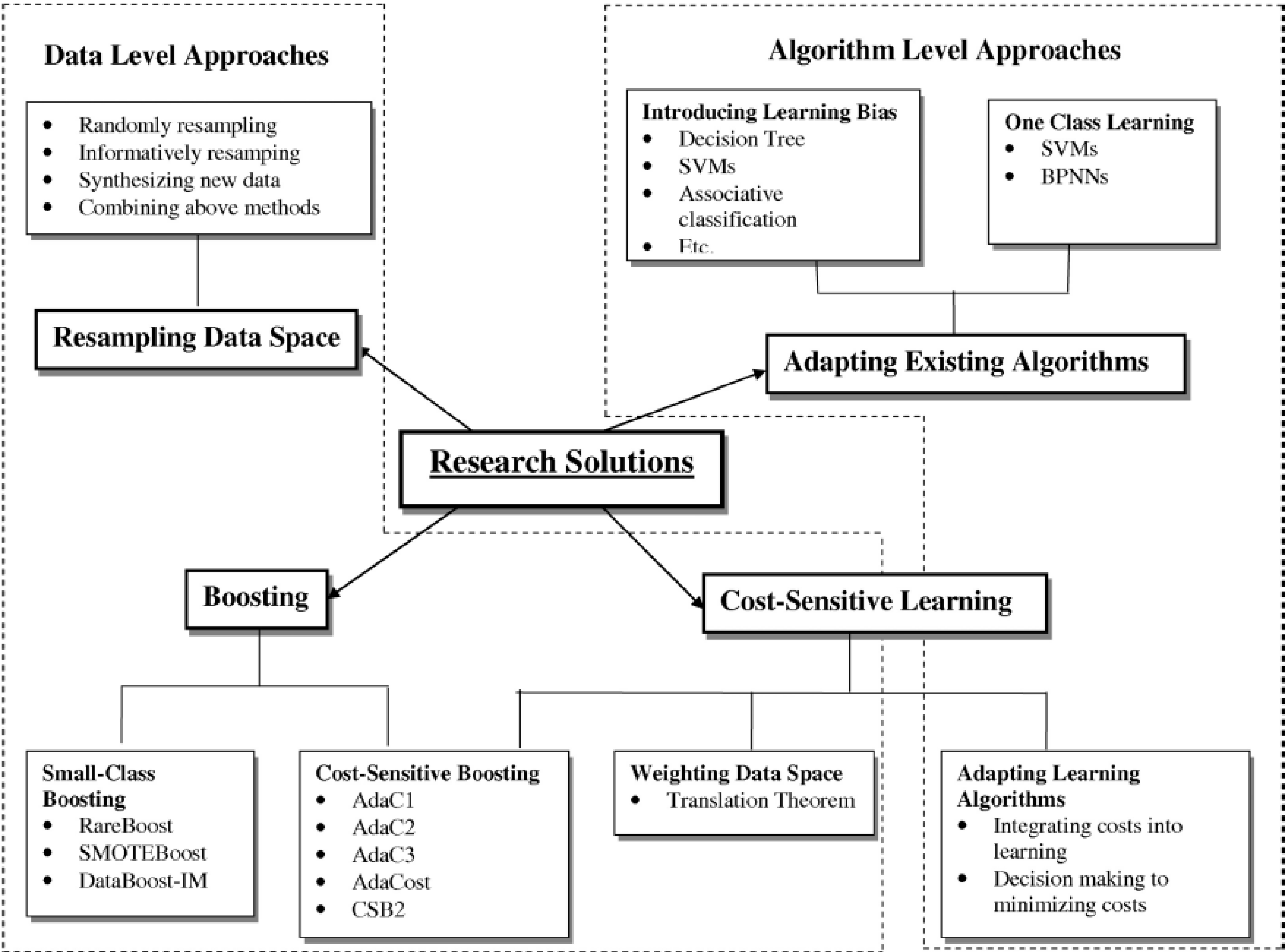
We divide our Imbalanced data sets into the following sections:

- Imbalance ratio between 1.5 and 9
- Imbalance ratio higher than 9 - Part I
- Imbalance ratio higher than 9 - Part II
- Multiple class imbalanced problems
- Noisy and Borderline Examples

We also include the preprocessed data sets.

Contents

Class Imbalance: Trends and final comments Data level vs algorithm Level



Y. Sun, A. K. C. Wong and M. S. Kamel. Classification of imbalanced data: A review. International Journal of Pattern Recognition 23:4 (2009) 687-719.

Class Imbalance: Trends and final comments New studies, trends and challenges

- Improvements on resampling – specialized resampling
    - New approches for creating artificial instances
    - How to choose the amount to sample?
    - New hybrid approaches oversampling vs undersampling
- Cooperation between resampling/cost sensitive/boosting
- Cooperation between feature selection and resampling
- Scalability: high number of features and sparse data
- Intrinsic data characteristics. To analyze the challenges on the class distribution.

Class Imbalance: Trends and final comments New studies, trends and challenges

In short, it is necessary to do work for:

Establishing some fundamental results regarding:

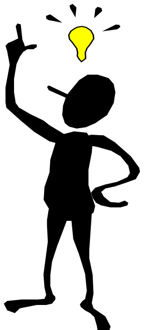
a) **the nature of the problem,**
b) **the behaviour of different types of classifiers, and**
c) **the relative performance of various previously proposed schemes for dealing with the problem.**

Designing new methods addressing the problem.
Tackling data preprocessing and changing rule classification strategy.

Class Imbalance: Trends and final comments Final comments

Class imbalance is a challenging and critical problem in the knowledge discovery field, the classification with imbalanced data sets.

Due to the intriguing topics and tremendous potential applications, the classification of imbalanced data will continue to receive more and more attention along next years. Class of interest is often much smaller or rarer (minority class).



Inteligencia de Negocio

TEMA 7. Modelos Avanzados de Minería de Datos

1. Clases no balanceadas/equilibradas

2. Características intrínsecas de los datos en clasificación

3. Detección de anomalías

4. Problemas no estándar de clasificación: MIL, MLL, …

5. Análisis de Sentimientos

6. Deep Learning

INTELIGENCIA DE NEGOCIO
2019 - 2020

Inteligencia de Negocio

- **Tema 1. Introducción a la Inteligencia de Negocio**
- **Tema 2. Minería de Datos. Ciencia de Datos**
- **Tema 3. Modelos de Predicción: Clasificación, regresión y series temporales**
- **Tema 4. Preparación de Datos**
- **Tema 5. Modelos de Agrupamiento o Segmentación**
- **Tema 6. Modelos de Asociación**
- **Tema 7. Modelos Avanzados de Minería de Datos**
- **Tema 8. Big Data**