I Workshop on Knowledge Extraction based on Evolutionary Learning
15-16 May, 2008

Introduction to Imbalanced data sets.

Some results on the use of evolutionary prototype selection for imbalanced data sets.

Salvador García López

In collaboration with F. Herrera

Research Group "Soft Computing and Intelligent Information Systems"

Department of Computer Science and Artificial Intelligence

University of Granada, 18071 – SPAIN

salvagl@decsai.ugr.es

http://sci2s.ugr.es

Introduction to Imbalanced Datasets

Learning in non-Balanced domains.

Data balancing through resampling.

State-of-the-art algorithm: *SMOTE.*

Introduction to Imbalanced Datasets

Learning in non-Balanced domains.

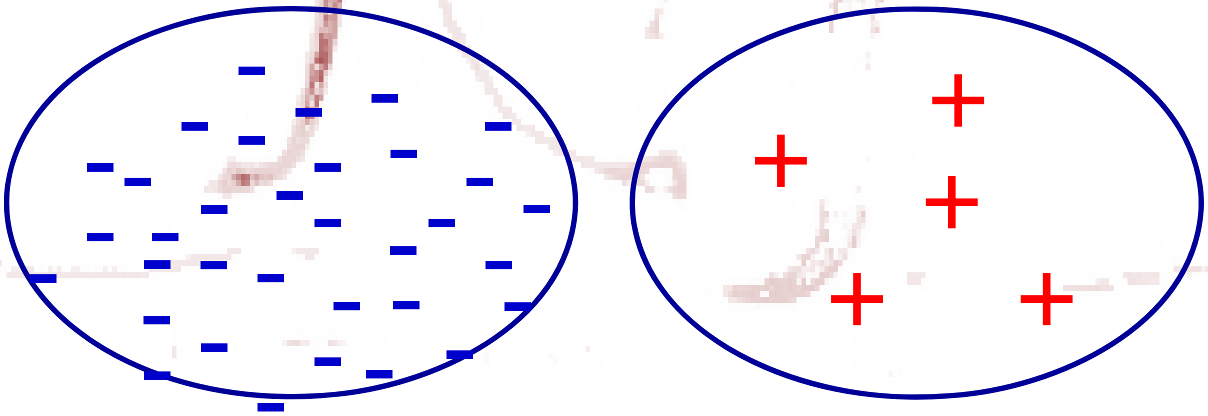
Data balancing through resampling.

State-of-the-art algorithm: *SMOTE.*

Learning in non-balanced domains

Data sets are said to be balanced if there are, approximately, as many positive examples of the concept as there are negative ones.

The positive examples are more interesting or their misclassification has a higher associate cost.



G. Cohen, M. Hilario, H. Sax, S. Hugonnet, A. Geisbuhler. Learning from Imbalanced Data in Surveillance of Nosocomial Infection. Artificial Intelligence in Medicine 37 (2006) 7-18

Learning in non-balanced domains

The classes of small size are usually labeled by rare cases (rarities).

The most important knowledge usually resides in the rare cases.

These cases are common in classification problems:

Ej.: Detection of uncommon diseases.

Imbalanced data: Few sick persons and lots of healthy persons.

Some real-problems:

- **Fraudulent credit card transactions**
- **Learning word pronunciation**
- **Prediction of pre-term births**
- **Prediction of telecommunications equipment failures**
- **Detection oil spills from satellite images**
- **Detection of Melanomas**

Learning in non-balanced domains

Problem:

- The problem with class imbalances is that standard learners are often biased towards the majority class.
- That is because these classifiers attempt to reduce global quantities such as the error rate, not taking the data distribution into consideration.

Result:

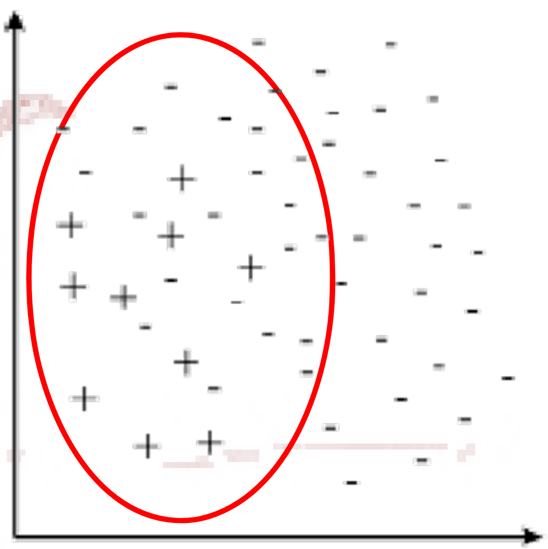
- As a result examples from the overwhelming class are well-classified whereas examples from the minority class tend to be misclassified.

Learning in non-balanced domains

¿Why is difficult to learn in imbalanced domains?

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

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



N.V. Chawla, N. Japkowicz, A. Kolcz. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6:1 (2004) 1-6

Learning in non-balanced domains

¿How can we evaluate an algorithm in imbalanced domains?

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

Confusion matrix for a two-class problem

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

Classical evaluation:

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

Learning in non-balanced domains

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

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

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

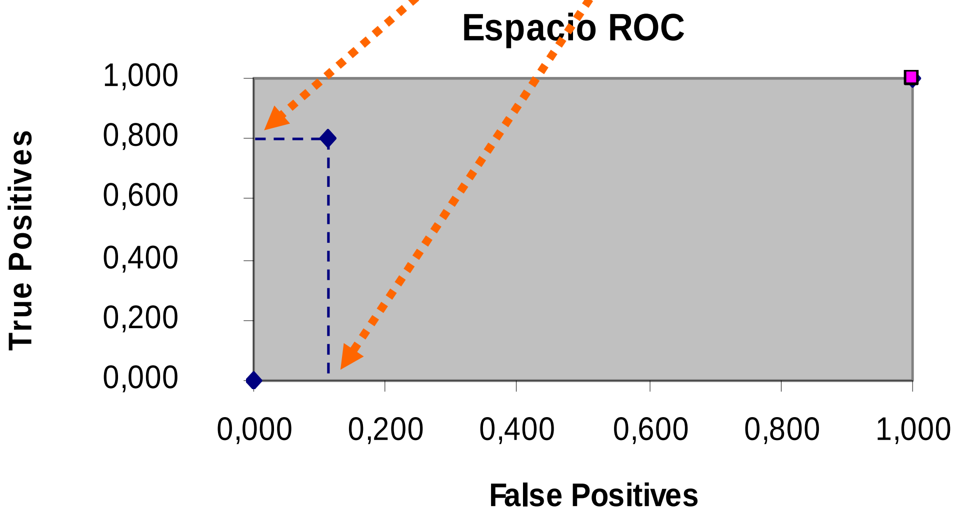
Learning in non-balanced domains

ROC Curves

The confusion matrix is normalized by columns

Real

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



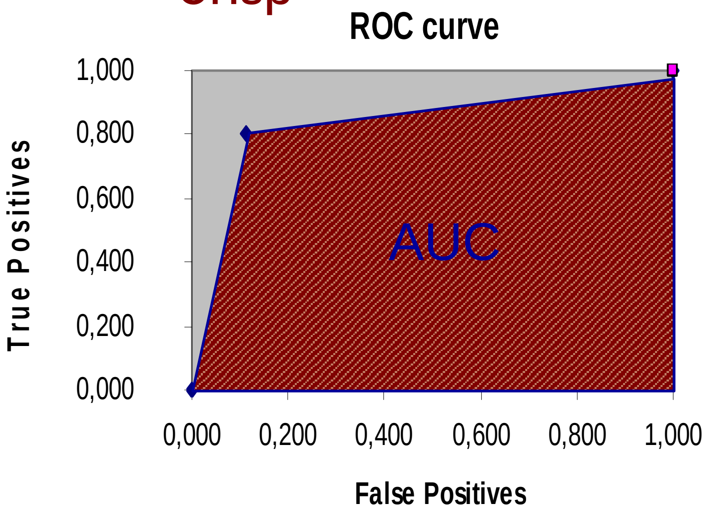
A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30(7) (1997) 1145-1159.

Learning in non-balanced domains

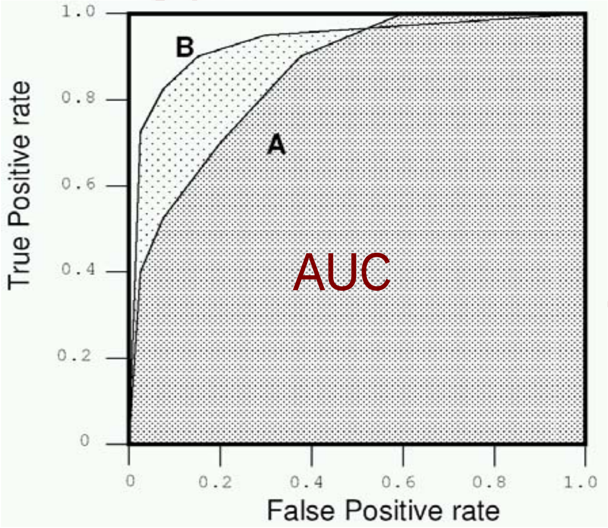
"crisp" and "soft" classifiers:

- **A "crisp" classifier (discrete) predicts a class among the candidates.**
- **A "soft" classifier (probabilistic) predicts a class, but this prediction is accompanied by a reliability value.**

Crisp



Soft



AUC: Área under ROC curve. Scalar quantity widle used for estimating classifiers performance.

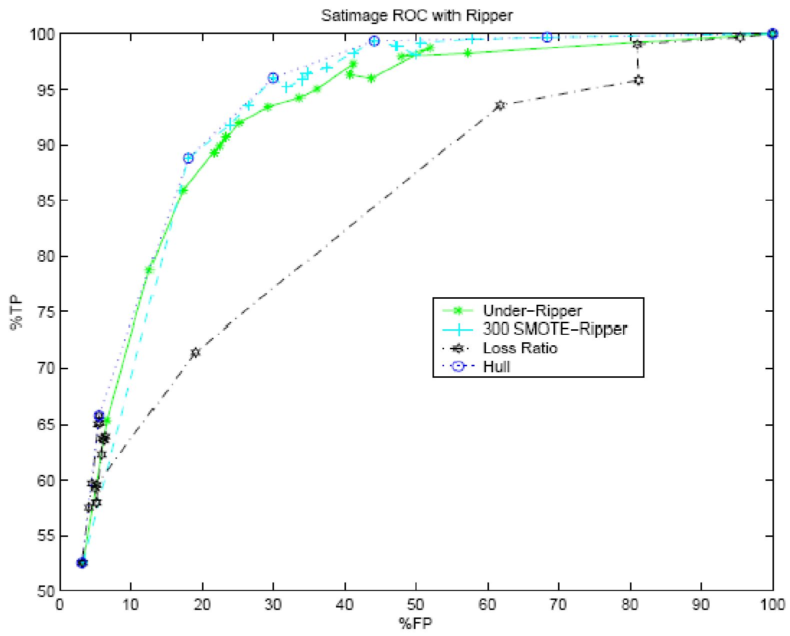
Learning in non-balanced domains

ROC analysis oriented to data resampling in imbalanced domains

The resampling algorithm must allow to adjust the rate of under/over sampling.

Performance of the classifier is measured with *over/under Sampling* at 25%, 50%, 100%, 200%, 300%, etc.

It can be only used in resmapling techniques which allow the adjustment of this parameter.



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

Introduction to Imbalanced Datasets

Learning in non-Balanced domains.

Data balancing through resampling.

State-of-the-art algorithm: *SMOTE.*

Data Balancing through *re-sampling*

Strategies

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

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

Cost Modifying

Motivation
- **Retain influyent examples**
- **Balance the training set**
- **Remove noisy instances in the decision boundaries**
- **Reduce the training set**

Data Balancing through *re-sampling*

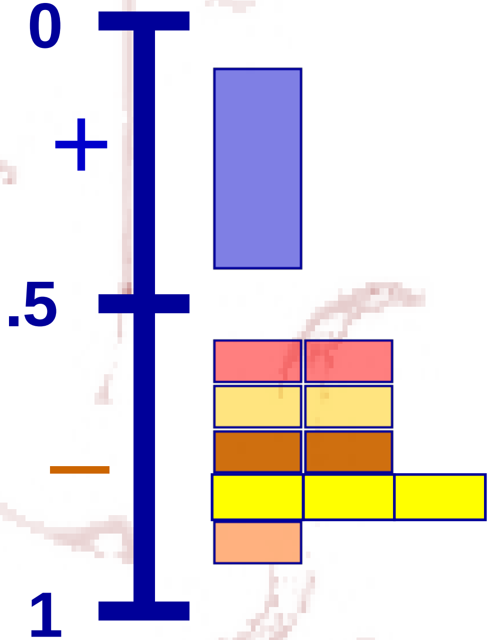
Over Sampling

- **Random**
- **Focused**

Under Sampling

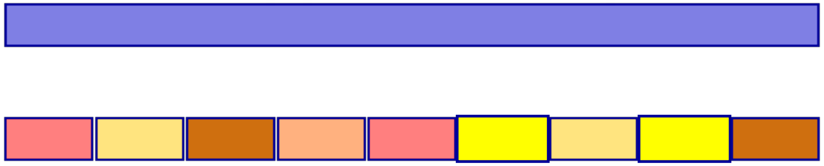
- **Random**
- **Focused**

Cost Modifying



examples of +

examples of –



Data Balancing through *re-sampling*

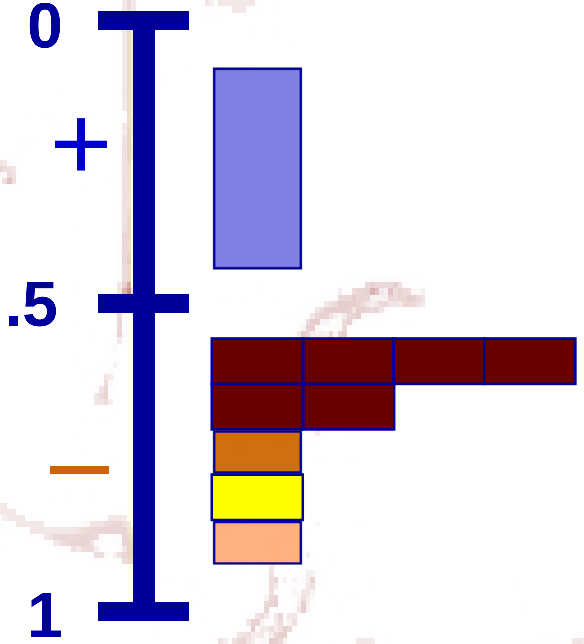
Over Sampling

- Random
- **Focused**

Under Sampling

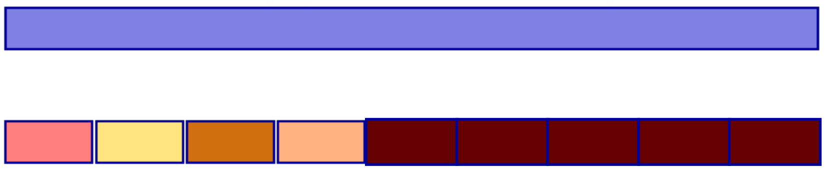
- Random
- Focused

Cost Modifying



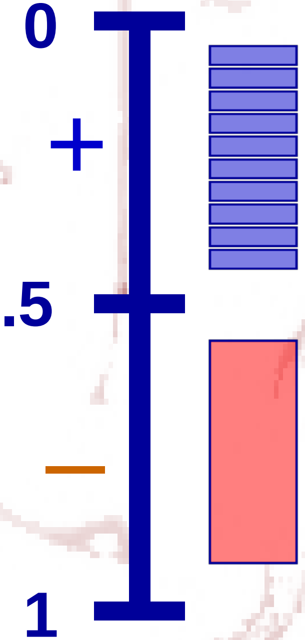
examples of +

examples of –



Data Balancing through *re-sampling*

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

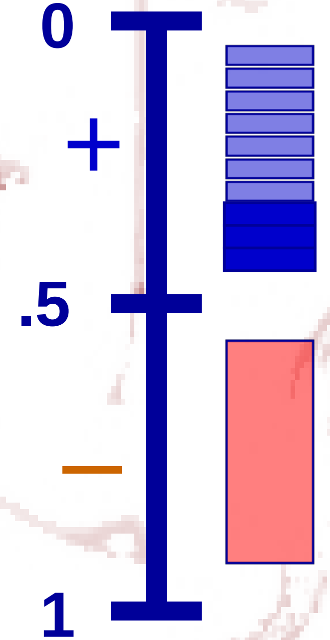


examples of + 

examples of – 

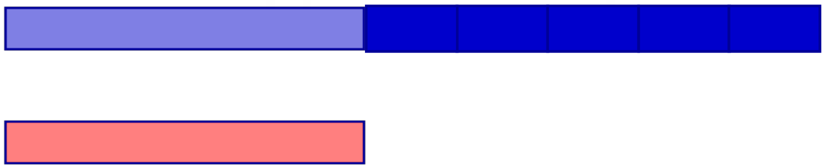
Data Balancing through *re-sampling*

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



\# examples of +

\# examples of –



Data Balancing through *re-sampling*

Over Sampling

Random

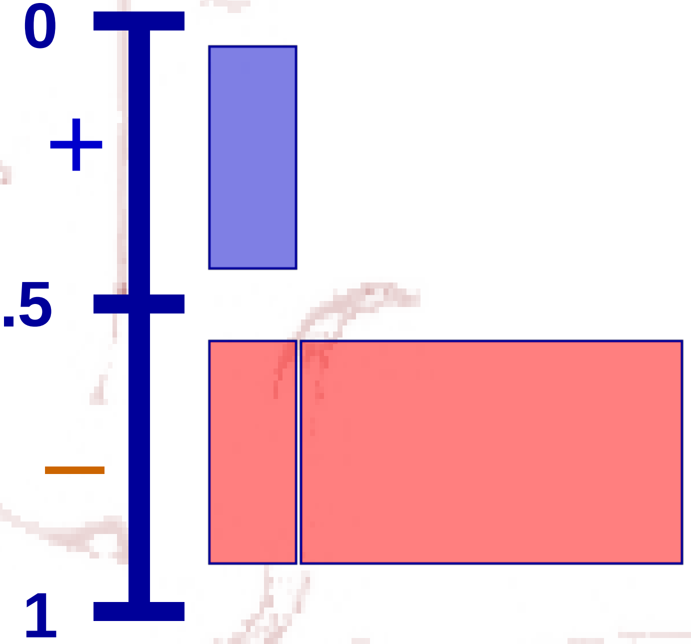
Focused

Under Sampling

Random

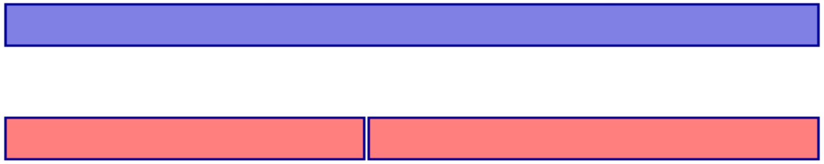
Focused

Cost Modifying



examples of +

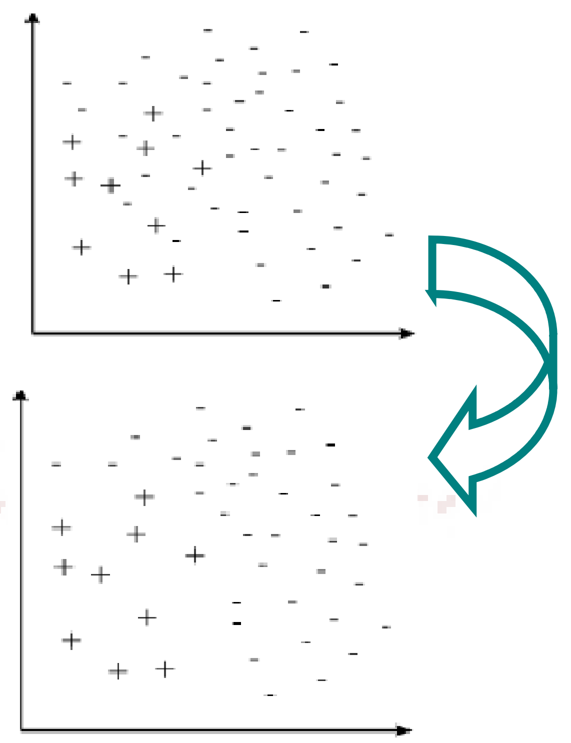
examples of –



Data Balancing through *re-sampling*

Under-sampling: Tomek Links

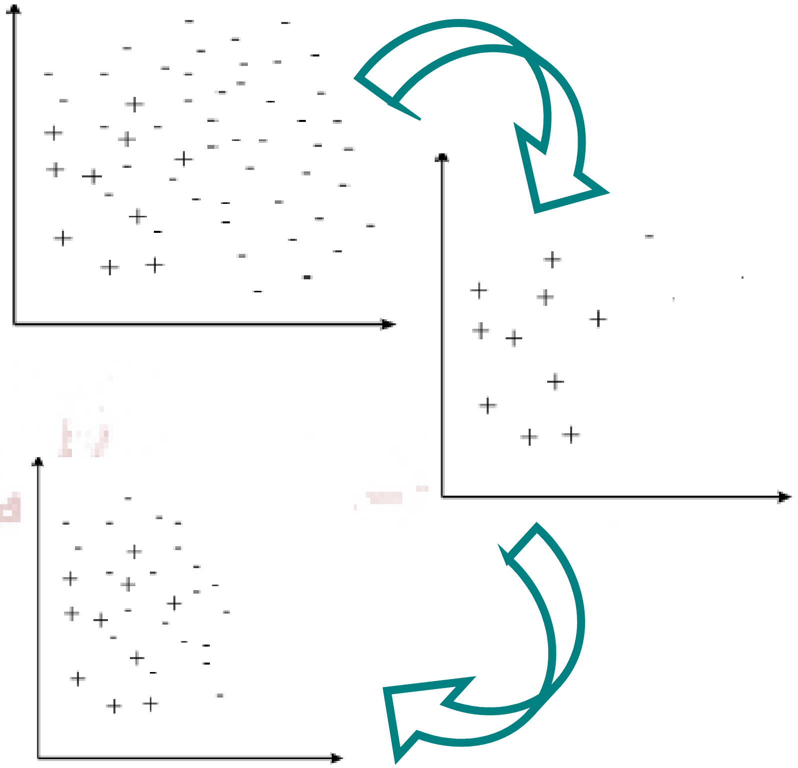
- To remove both noise and borderline examples
- Tomek link
  - –E_i, E_j belong to different classes, d (E_i, E_j) is the distance between them.
  - –A (E_i, E_j) pair is called a Tomek link if there is no example E_l, such that $d(E_i, E_l) < d(E_i, E_j)$ or $d(E_j , E_l) < d(E_i, E_j)$.



Data Balancing through *re-sampling*

Under-sampling: US-CNN

- To remove both noise and borderline examples
- Algorithm:
  - Let E be the original training set
  - Let E' contains all positive examples from S and one randomly selected negative example
  - Classify E with the 1-NN rule using the examples in E'
  - Move all misclassified example from E to E'



Data Balancing through *re-sampling*

Under-sampling:

- One-sided selection
  - –Tomek links + CNN
- CNN + Tomek links
  - –Proposed by the author
  - –Finding Tomek links is computationally demanding, it would be computationally cheaper if it was performed on a reduced data set.

- NCL
  - To remove majority class examples
  - Different from OSS, emphasize more data cleaning than data reduction
  - Algorithm:
    - –Find three nearest neighbors for each example E_i in the training set
    - –If E_i belongs to majority class, & the three nearest neighbors classify it to be minority class, then remove E_i
    - –If E_i belongs to minority class, and the three nearest neighbors classify it to be majority class, then remove the three nearest neighbors

Introduction to Imbalanced Datasets

Learning in non-Balanced domains.

Data balancing through resampling.

State-of-the-art algorithm: *SMOTE.*

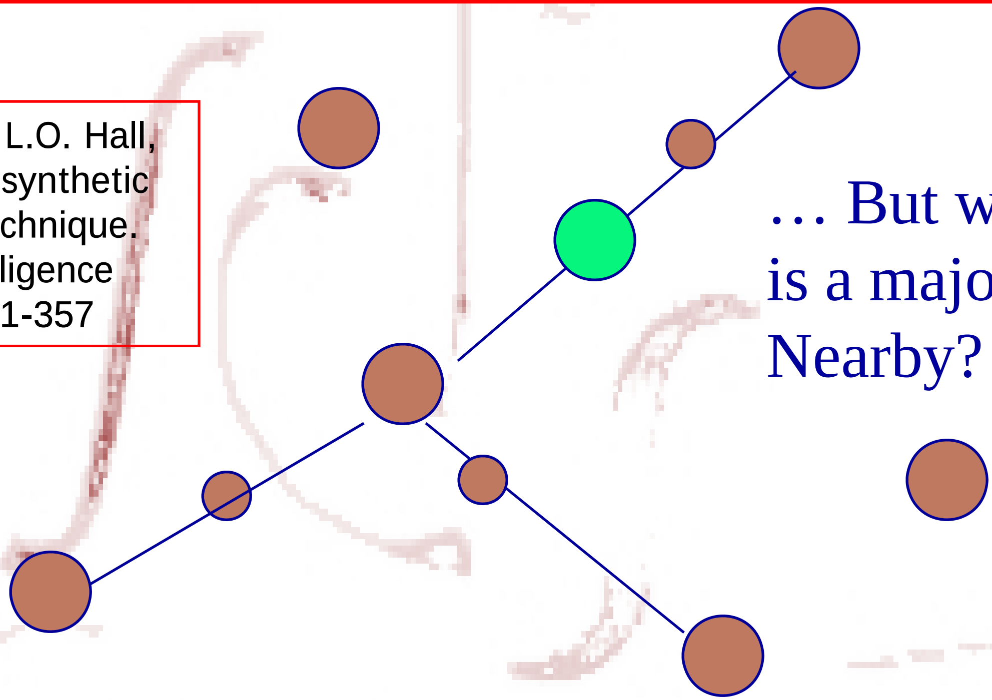
State-of-the-art algorithm: SMOTE.

Over-sampling method:

- To form new minority class examples by interpolating between several minority class examples that lie together.
- in ``feature space'' rather than ``data space''
- Algorithm: For each minority class example, introduce synthetic examples along the line segments joining any/all of the k minority class nearest neighbors.
- Note: Depending upon the amount of over-sampling required, neighbors from the *k* nearest neighbors are randomly chosen.
- For example: if we are using 5 nearest neighbors, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each.

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



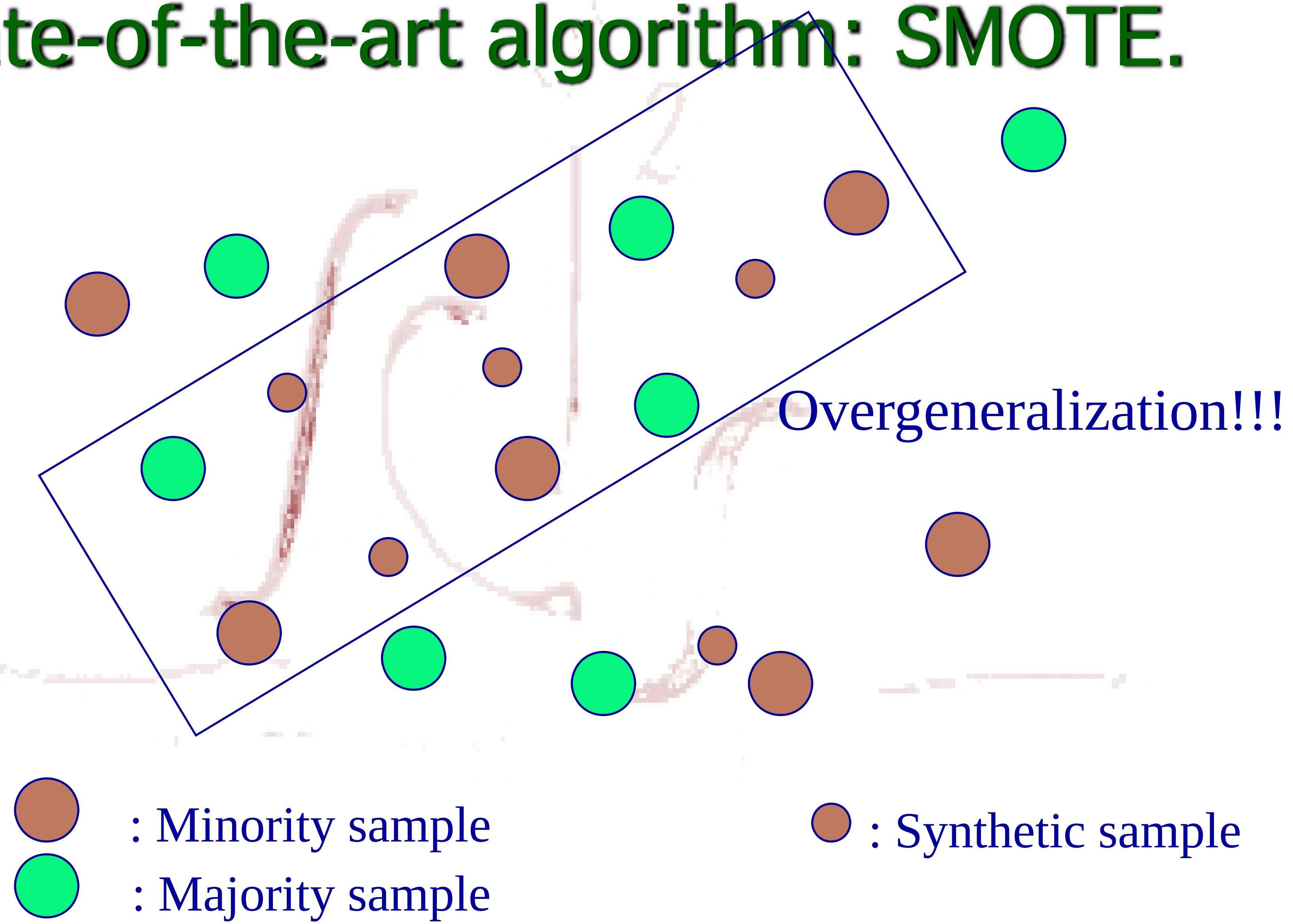
… But what if there is a majority sample Nearby?

: Minority sample

: Synthetic sample

: Majority sample

State-of-the-art algorithm: SMOTE.

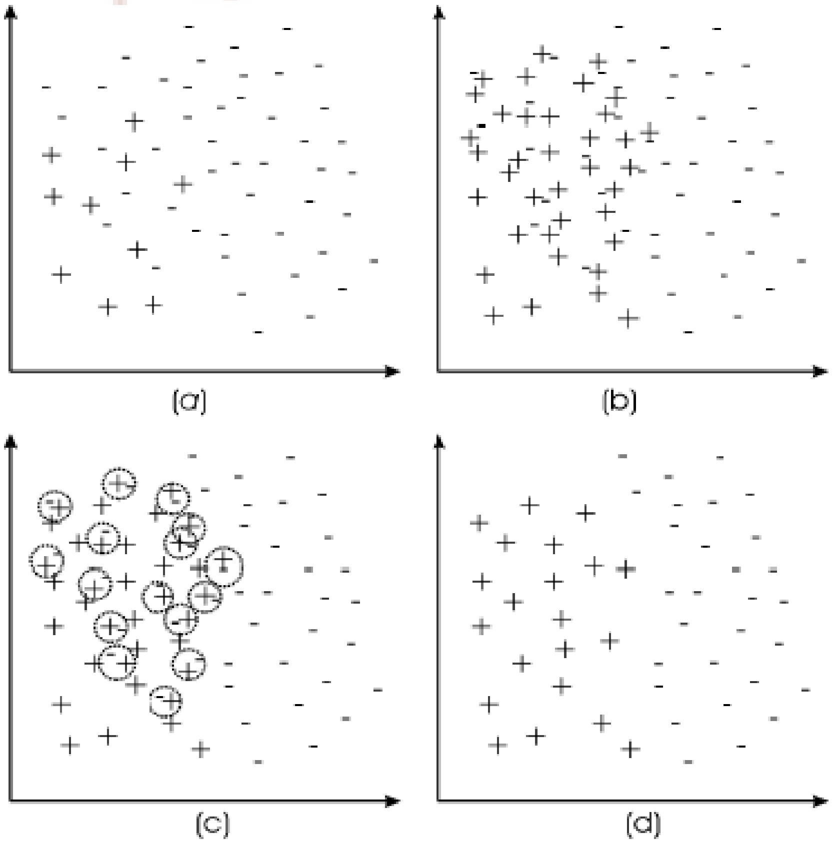


State-of-the-art algorithm: SMOTE.

SMOTE

+

TomekLinks



State-of-the-art algorithm: SMOTE.

SMOTE + ENN:

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

State-of-the-art algorithm: SMOTE.

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

Data set	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º
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

Some results on the use of evolutionary prototype selection for imbalanced data sets

Evolutionary Under-Sampling

Experimental Framework and Results

Conclusions and Future Work

Some results on the use of evolutionary prototype selection for imbalanced data sets

Evolutionary Under-Sampling

Experimental Framework and Results

Conclusions and Future Work

Evolutionary Under-Sampling

Evolutionary algorithm for re-sampling:

Representation:

0	1	1	1	0	0	1	0	0	1

Base Method: CHC

It introduces different features to obtain a trade-off between exploration and exploitation; such as incest prevention, reinitialization of the search process when it becomes blocked and the competition among parents and offspring into the replacement process

Models:

- **EBUS:** Aim for an optimal balancing of data without loss of effectiveness in classification accuracy
- **EUSCM:** Aim for an optimal power of classification without taking into account the balancing of data, considering the latter as a subobjective that may be an implicit process.

Evolutionary Under-Sampling

Type of Selection:

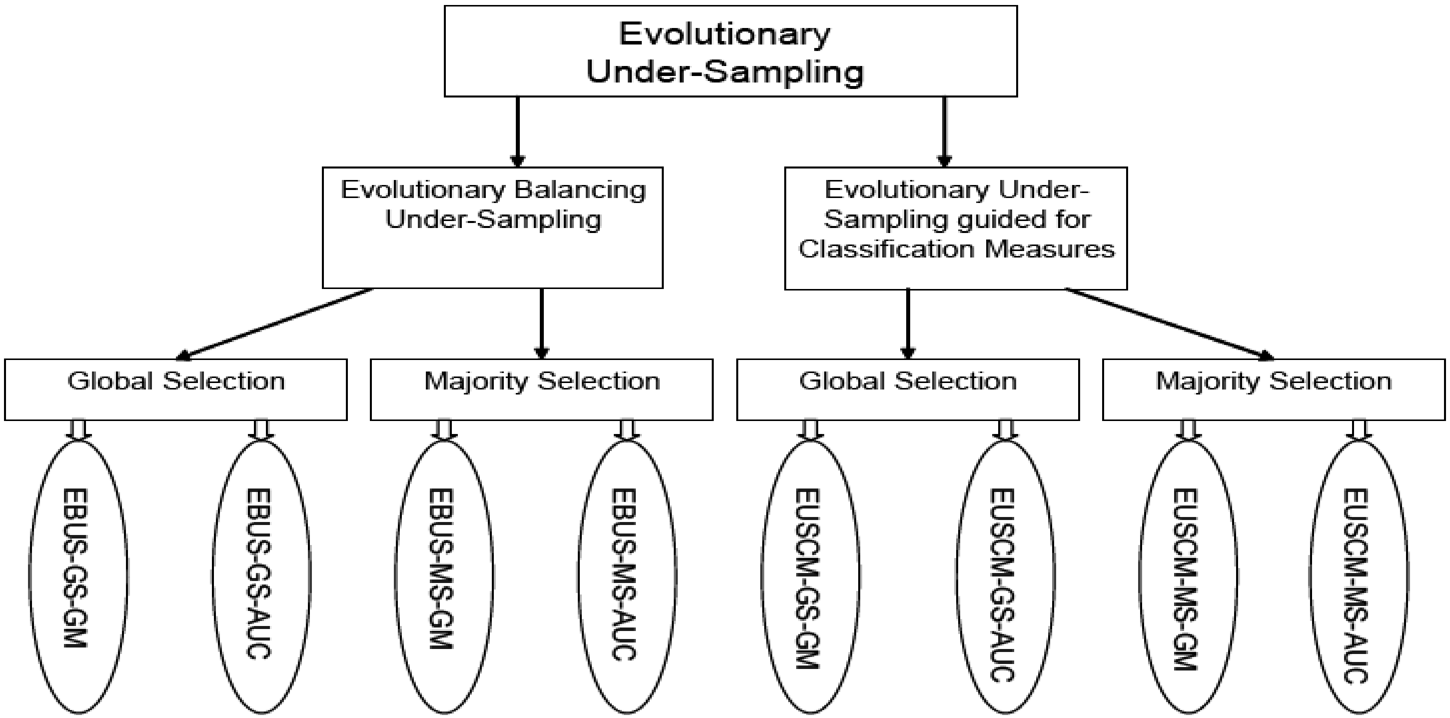
- GS: **Global Selection,** the selection scheme proceeds over any kind of instance.
- MS: **Majority Selection,** the selection scheme only proceeds over majority class instances.

Evaluation Measures:

- GM: **Geometric Mean**
- AUC: **Area under ROC Curve**

Evolutionary Under-Sampling

Taxonomy:



Evolutionary Under-Sampling

Fitness function in EBUS model:

$$Fitness_{Bal}(S) = \begin{cases} g - |1 - \frac{n^+}{n^-}| \cdot P & \text{if } n^- > 0 \\ g - P & \text{if } n^- = 0 \end{cases} \qquad Fitness_{Bal}(S) = \begin{cases} AUC - |1 - \frac{n^+}{n^-}| \cdot P & \text{if } n^- > 0 \\ AUC - P & \text{if } n^- = 0 \end{cases}$$

P: **is a penalization factor that controls the intensity and importance of the balance during the evolutionary search.**

P = 0.2 works appropriately.

Fitness function in EUSCM model:

$$Fitness(S) = g, \qquad Fitness(S) = AUC,$$

Some results on the use of evolutionary prototype selection for imbalanced data sets

Evolutionary Under-Sampling

Experimental Framework and Results

Conclusions and Future Work

Experimental Framework and Results

Algorithms used in the comparison:

Prototype Selection:

IB3 DROP3 EPS-CHC EPS-IGA

Undersampling:

Random Under-Samplig TomekLinks (TL)

CNN OSS CNN+TL NCL

CPM SBC

Under-Sampling based on clustering

Experimental Framework and Results

Data sets:

IR:

Imbalance ratio:

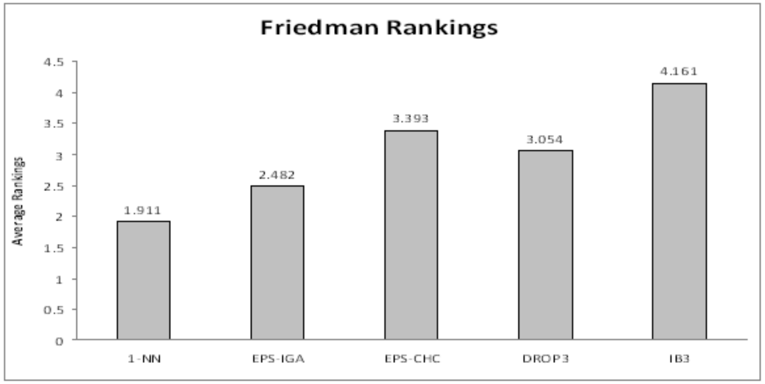
Number negative examples / Number positive examples

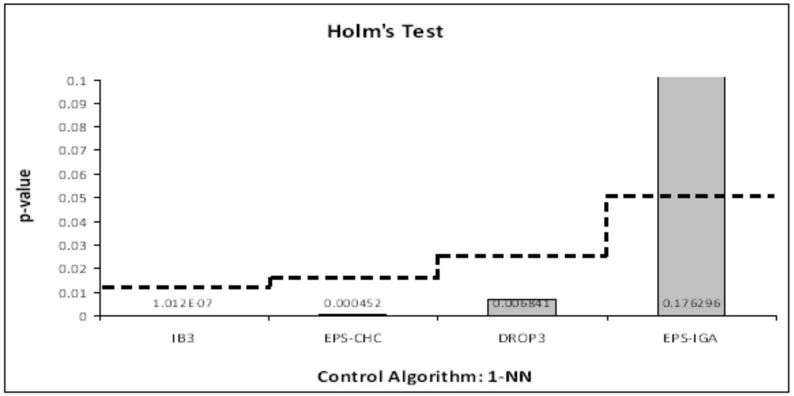
Data set	#Examples	#Attributes	Class (min., maj.)	%Class(min.,maj.)	IR
GlassBWNFP	214	9	(build-window-non_float-proc, remainder)	(35.51, 64.49)	1.82
EcoliCP-IM	220	7	(im,cp)	(35.00, 65.00)	1.86
Pima	768	8	(1,0)	(34.77, 66.23)	1.9
GlassBWFP	214	9	(build-window-float-proc, remainder)	(32.71, 67.29)	2.06
German	1000	20	(1, 0)	(30.00, 70.00)	2.33
Haberman	306	3	(Die, Survive)	(26.47, 73.53)	2.68
Splice-ie	3176	60	(ie,remainder)	(24.09, 75.91)	3.15
Splice-ei	3176	60	(ei,remainder)	(23.99, 76.01)	3.17
GlassNW	214	9	(non-windows glass, remainder)	(23.93, 76.17)	3.19
VehicleVAN	846	18	(van,remainder)	(23.52, 76.48)	3.25
EcoliIM	336	7	(im,remainder)	(22.92, 77.08)	3.36
New-thyroid	215	5	(hypo,remainder)	(16.28, 83.72)	4.92
Segment1	2310	19	(1,remainder)	(14.29, 85.71)	6.00
EcoliIMU	336	7	(iMU, remainder)	(10.42, 89.58)	8.19
Optdigits0	5564	64	(0, remainder)	(9.90, 90.10)	9.10
Satimage4	6435	36	(4, remainder)	(9.73, 90.27)	9.28
Vowel0	990	13	(0, remainder)	(9.01, 90.99)	10.1
GlassVWFP	214	9	(Ve-win-float-proc, remainder)	(7.94, 92.06)	10.39
EcoliOM	336	7	(om, remainder)	(6.74, 93.26)	13.84
GlassContainers	214	9	(containers, remainder)	(6.07, 93.93)	15.47
Abalone9-18	731	9	(18, 9)	(5.75, 94.25)	16.68
GlassTableware	214	9	(tableware, remainder)	(4.2, 95.8)	22.81
YeastCYT-POX	483	8	(POX, CYT)	(4.14, 95.86)	23.15
YeastME2	1484	8	(ME2, remainder)	(3.43, 96.57)	28.41
YeastME1	1484	8	(ME1, remainder)	(2.96, 97.04)	32.78
YeastEXC	1484	8	(EXC, remainder)	(2.49, 97.51)	39.16
Car	1728	6	(good, remainder)	(3.99, 96.01)	71.94
Abalone19	4177	9	(19, remainder)	(0.77, 99.23)	128.87

S. García – Some results on the use of evolutionary prototype selection for imbalanced data sets.

Experimental Framework and Results

Part I: Classical prototype selection as imbalanced undersampling

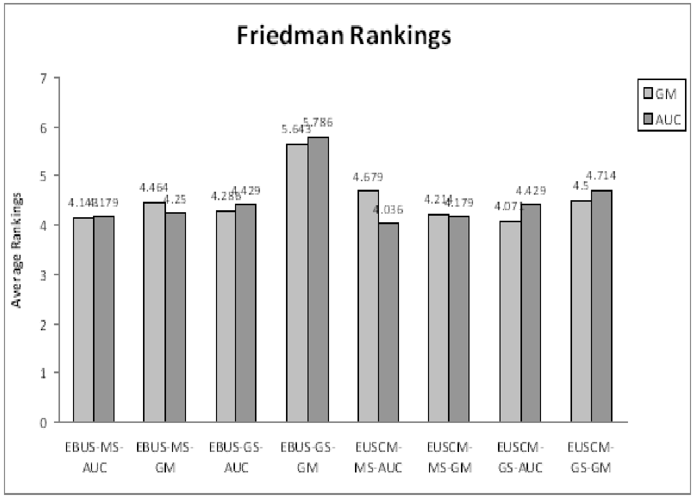


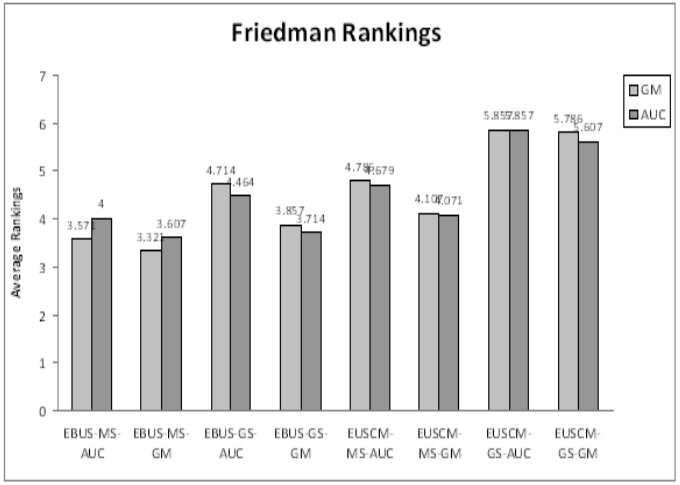


Classical prototype selection is not recommendable for tackling imbalanced data sets. 1-NN without preprocessing behaves the best.

Experimental Framework and Results

Part II: **Comparison among the eight proposals of Evolutionary Under-Sampling**





IR < 9 **IR > 9**

Experimental Framework and Results

Part II: **Comparison among the eight proposals of Evolutionary Under-Sampling**

IR < 9:

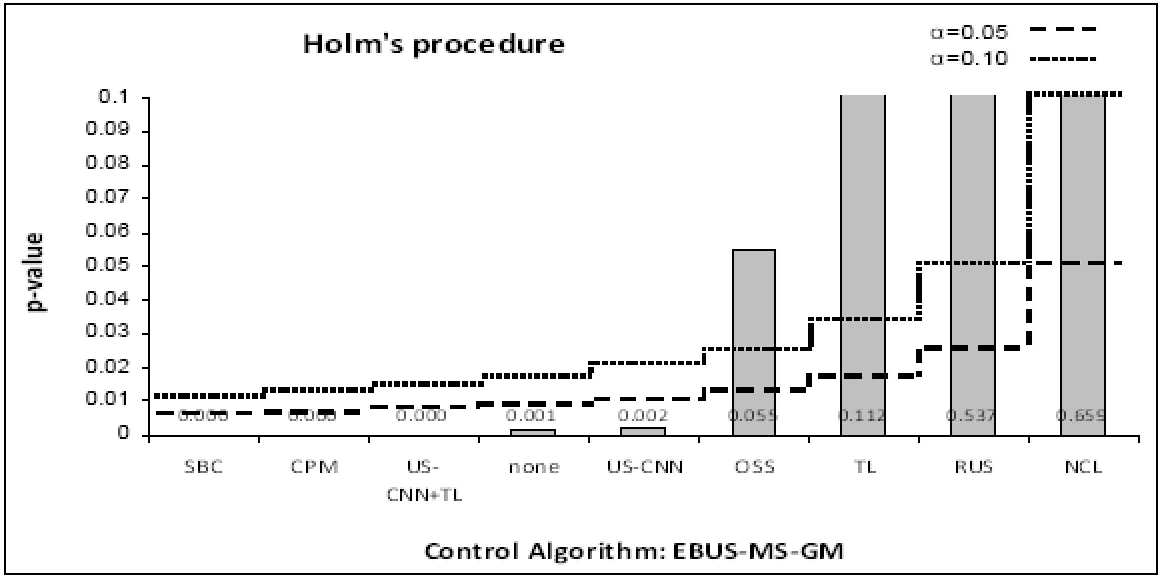
- **EUSCM behaves better than EBUS (P factor has little interest)**
- **Little differences between GM and AUC.**

IR > 9:

- **GS mechanism has no sense due to the high imbalance ratio. MS is preferable.**
- **P factor is very useful in this case. EBUS outperforms EUSCM**

Experimental Framework and Results

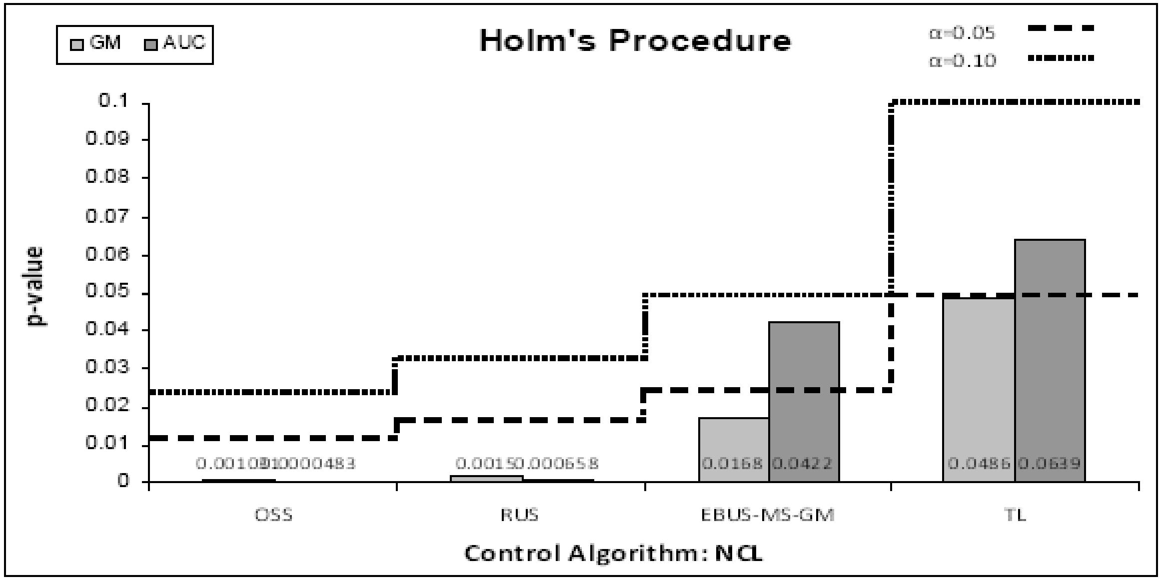
Part III: Comparison with other under-sampling approaches



Considering all data sets

Experimental Framework and Results

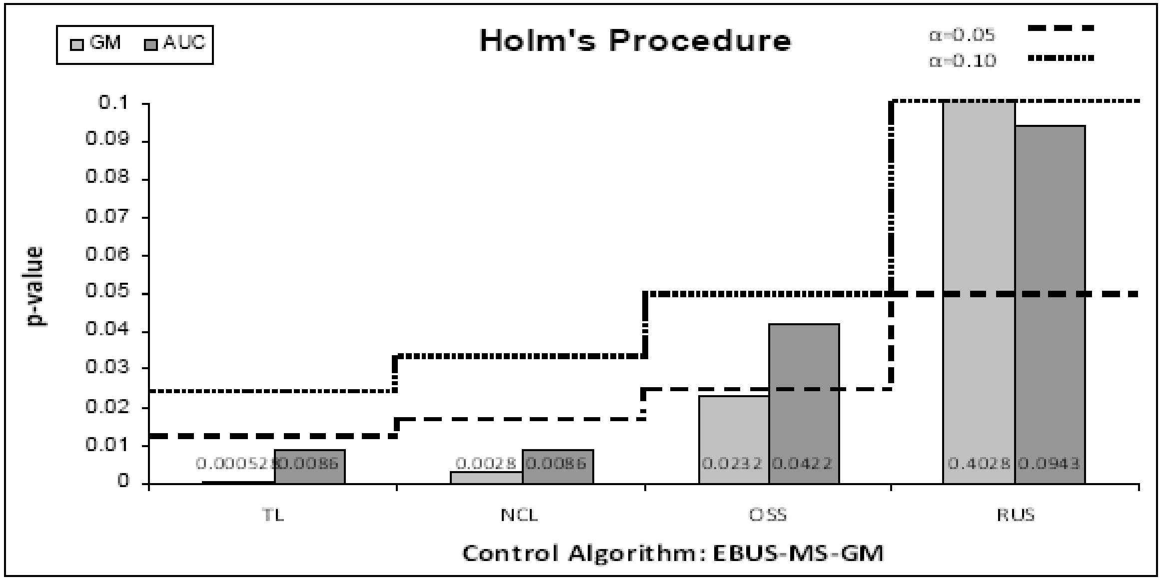
Part III: Comparison with other under-sampling approaches



Considering data sets with IR < 9

Experimental Framework and Results

Part III: Comparison with other under-sampling approaches



Considering data sets with IR > 9

Experimental Framework and Results

Part III: Comparison with other under-sampling approaches

- **EUS models usually present an equal or better performance than the remaining methods, independently of the degree of imbalance of data.**
- **The best performing under-sampling model over imbalance data sets is EBUS-MSGM**
- **The tendency of the EUS models follows an improving of the behaviour in classification when the data turns to a high degree of imbalance.**

Some results on the use of evolutionary prototype selection for imbalanced data sets

Evolutionary Under-Sampling

Experimental Framework and Results

Conclusions and Future Work

Conclusions and Future Work

- Prototype Selections methods are not useful when handling imbalanced problems.
- Evolutionary under-sampling is an effective model in instance-based learning.
- Majority selection mechanism obtains more accurate subsets of instances, but presents a lower reduction rate.
- No difference between GM and AUC (different evaluation measures) is observed.
- For dealing with low imbalance rates, EUSCM model is the best choice
- For dealing with high imbalance rates, EBUS model is the best.

Conclusions and Future Work

FUTURE WORK

- **Use of evolutionary under-sampling in training set selection, in order to optimize the performance of other classification algorithms.**
- **Study the scalability of these models in very large data sets.**
- **Hybridize evolutionary under-sampling with SMOTE or other over-sampling approaches.**
- **Analize the data in terms of data complexity in order to guide EUS to a better selection of instances and obtain generalized subsets.**

References

- Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6(1):20–29.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intel Res 16:321–357.
- Chawla NV, Japkowicz N, Kolcz A (2004) Editorial: learning from imbalanced datasets. SIGKDD Explorations 6(1):1–6
- Drummond C, Holte R (2003) C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In: Proceedings of the ICML'03 workshop on learning from imbalanced data sets
- Weiss G, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intel Res 19:315–354
- Chawla NV, Cieslak DA, Hall LO, Joshi A (2008) Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery. In press.
- García S, Herrera F (2008) Evolutionary Under-Sampling for Classification with Imbalanced Data Sets: Proposals and Taxonomy. Evolutionary Computation. In press.

Selección de Instancias y Extracción de Modelos. Dominios no Balanceados

Thanks!!!

DECSAI
Universidad de Granada