Tema – Artigo 1
(TP Intermediário)

Cristiano Leite de Castro - crislcastro@ufmg.br

Leitura (TP Intermediário)

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 861–874

Pattern Recognition Letters

www.elsevier.com/locate/patrec

An introduction to ROC analysis

Tom Fawcett

Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA

Available online 19 December 2005

Abstract

Receiver operating characteristics (ROC) graphs are useful for organizing classifiers and visualizing their performance. ROC graphs are commonly used in medical decision making, and in recent years have been used increasingly in machine learning and data mining research. Although ROC graphs are apparently simple, there are some common misconceptions and pitfalls when using them in practice. The purpose of this article is to serve as an introduction to ROC graphs and as a guide for using them in research.

ROC Analysis

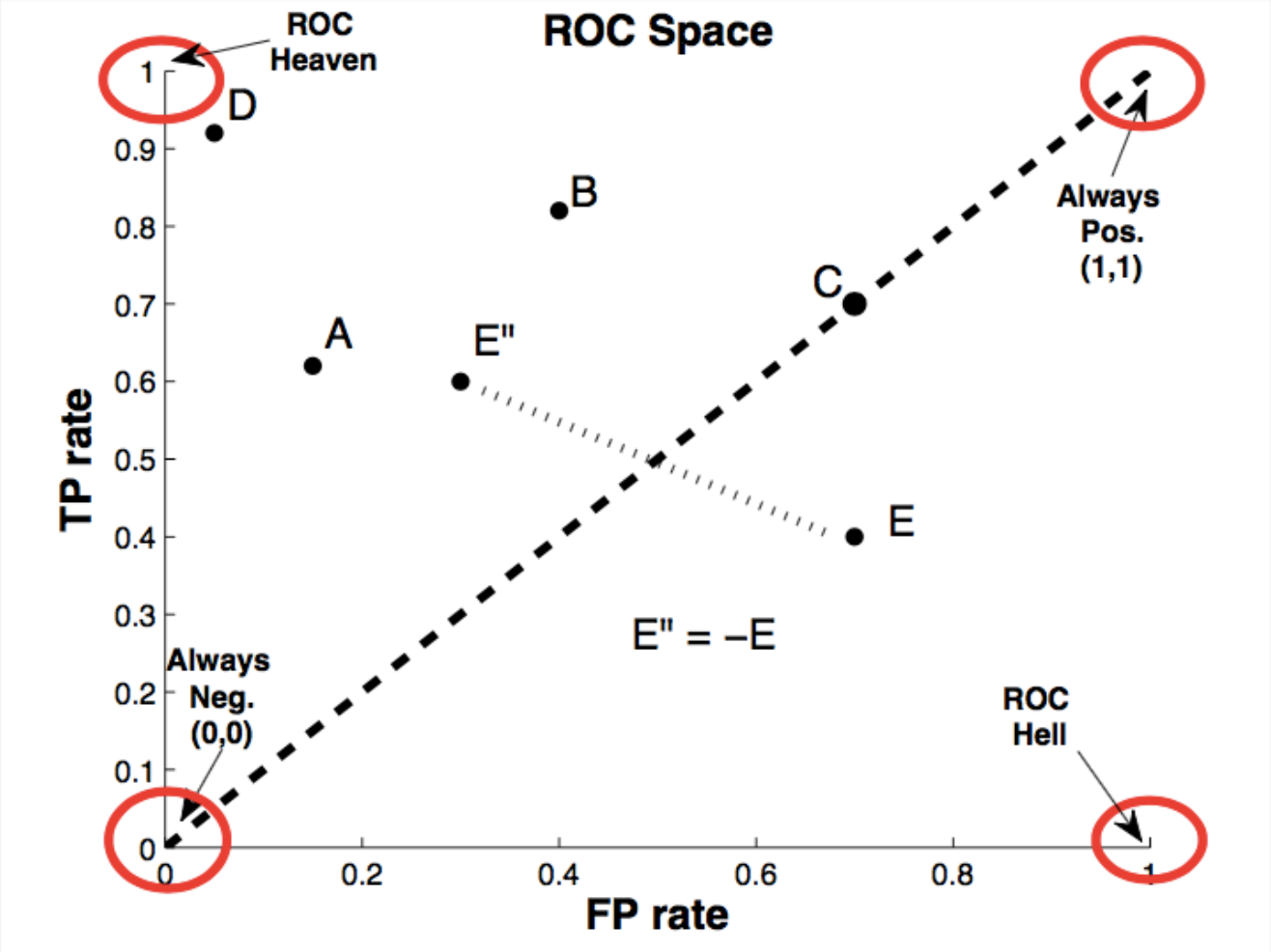
- $f : \mathbb{R}^n \rightarrow \{+1, -1\}$
- $C_1 \Rightarrow$ positive and $C_2 \Rightarrow$ negative.
- Basic idea: distinguish performance between classes.

Confusion Matrix

	predicted pos	predicted neg
actual pos	*TP*	*FN*
actual neg	*FP*	*TN*

ROC Analysis

ROC Graph



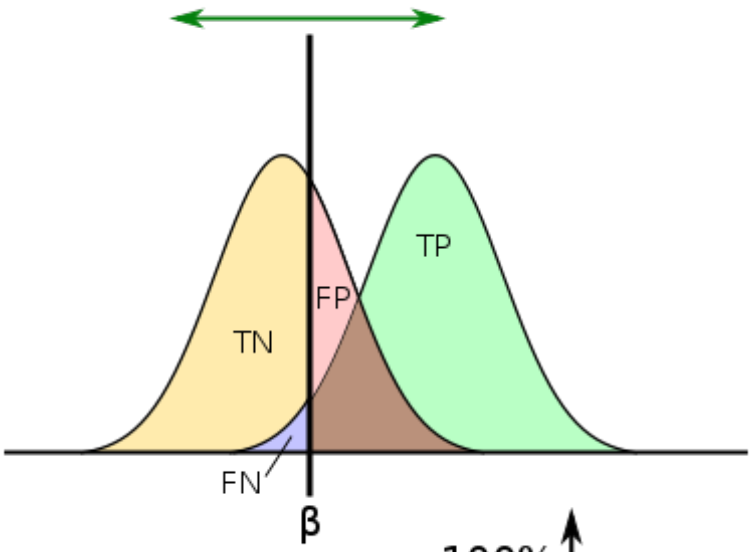
ROC Curves

- score $\hat{f}(\mathbf{x})$: refere-se à saída contínua do classificador;
  - indica pertinência (ou probabilidade) do padrão $\mathbf{x}$ pertencer à classe positiva.

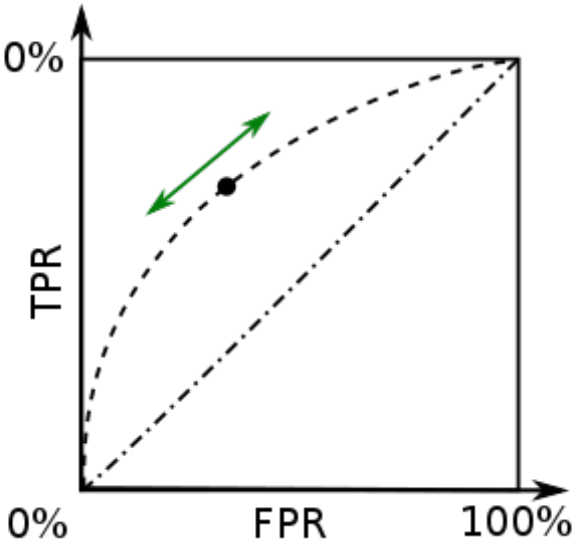
- ao se aplicar: $limiar(\hat{f}(\mathbf{x})) \rightarrow -1/1 \quad \forall\, \mathbf{x}$
  - gera um ponto (FPR, TPR) no ROC space

- Curva ROC:
  - varie o limiar sobre toda a faixa: $[min(\hat{f}(\mathbf{x})), max(\hat{f}(\mathbf{x}))]$ e plote (FPR, TPR) para cada limiar.

ROC Curves

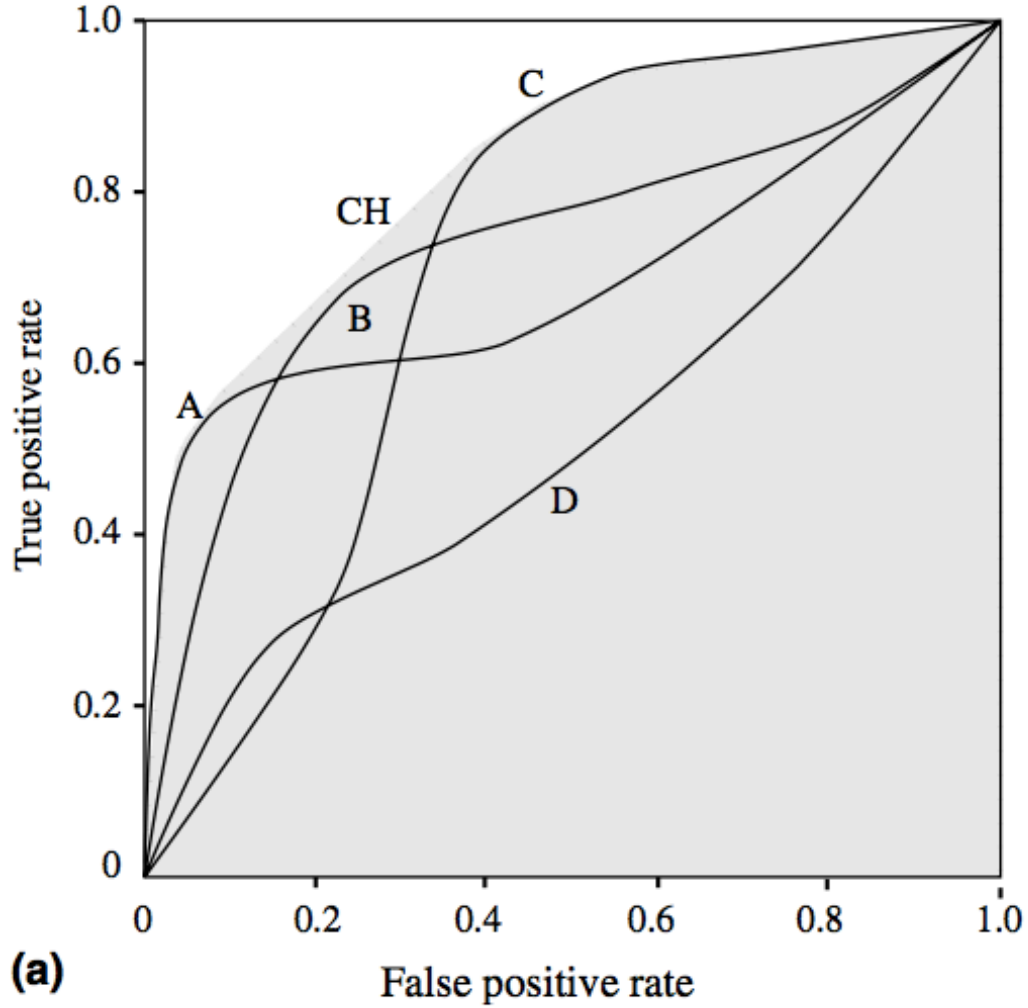


TP	FP
FN	TN



SEE ROC CURVE DEMONSTRATION:
Link: http://arogozhnikov.github.io/2015/10/05/roc-curve.html

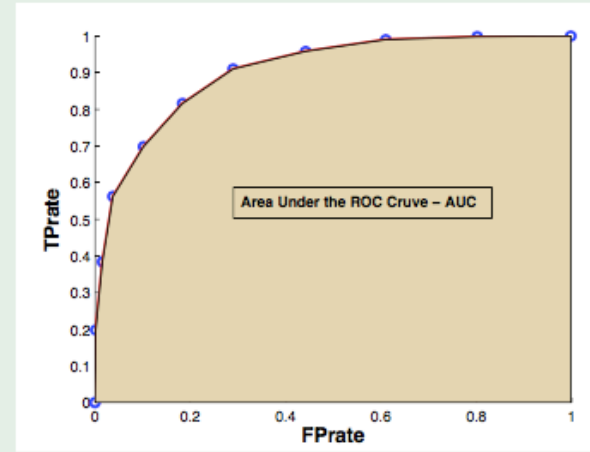
ROC Curves - Comparison



AUC (Area Under the ROC Curve)

Area Under the ROC Curve

- probability that randomly chosen $\mathbf{x}_j \in C_1$ is ranked higher than randomly chosen $\mathbf{x}_k \in C_2$.
- ranking quality.



Wilcoxon-Mann-Whitney

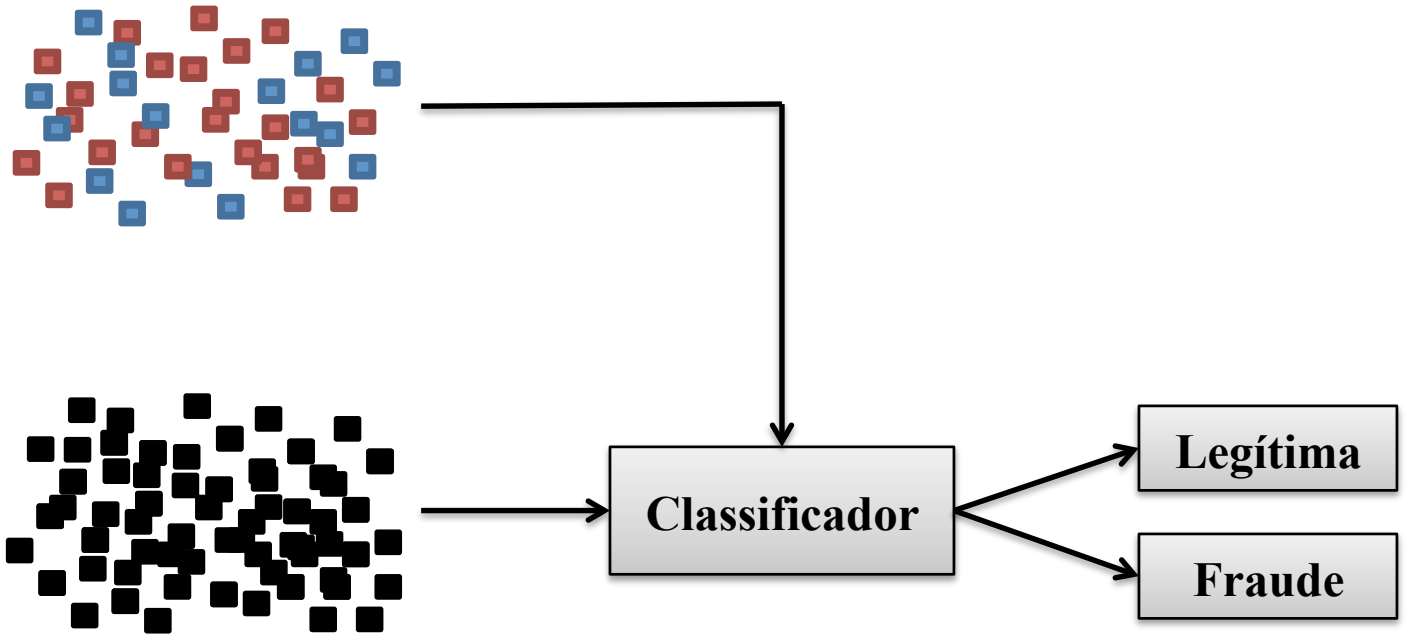
$$\widehat{AUC(f)} = \frac{1}{pn}\left(\sum_{j=1}^{p}\sum_{k=1}^{n} g(f(\vec{x_j^+}) - f(\vec{x_k^-}))\right)$$

$$g(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.5 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

PROBLEMA
(TEMA DO TP INTERMEDIÁRIO)

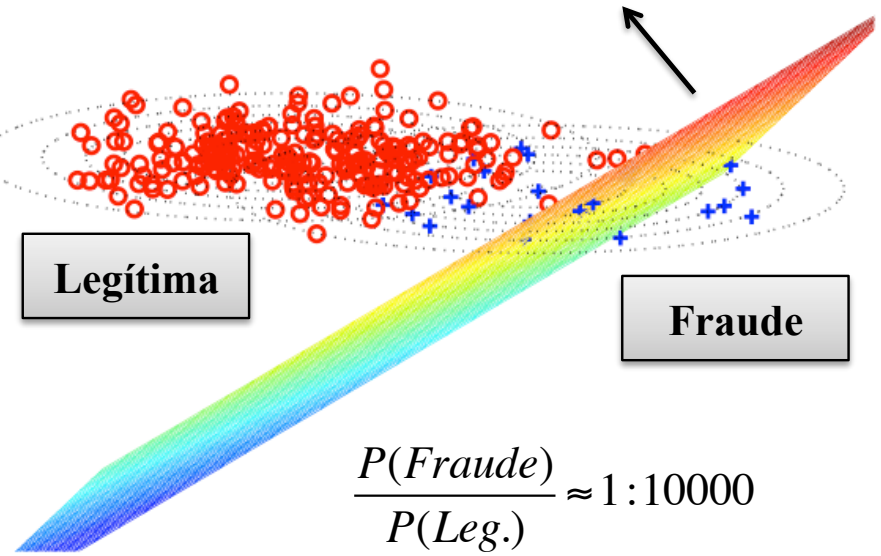
Contextualização do Problema

Classificação Binária (2 categorias)



Fonte: Provost and Fawcett. **Adaptive Fraud Detection**. Data Mining and Knowledge Discovery, 1997.

Problema de Classes Debalanceadas



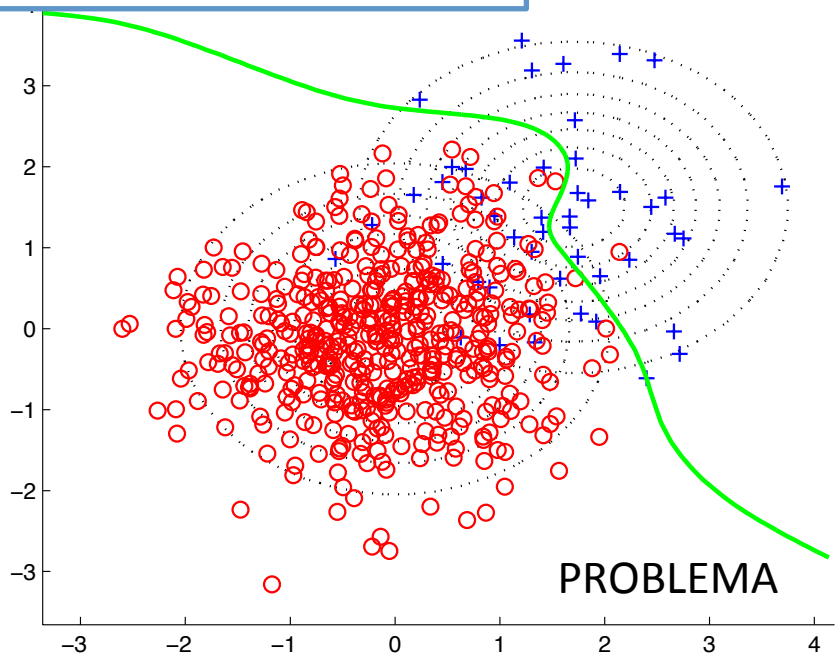
- **Problema aparece quando:**

- elevado desequilíbrio entre as classes;
- presença de incerteza (ruído);
- dados não são suficientes para representar a classe minoritária.

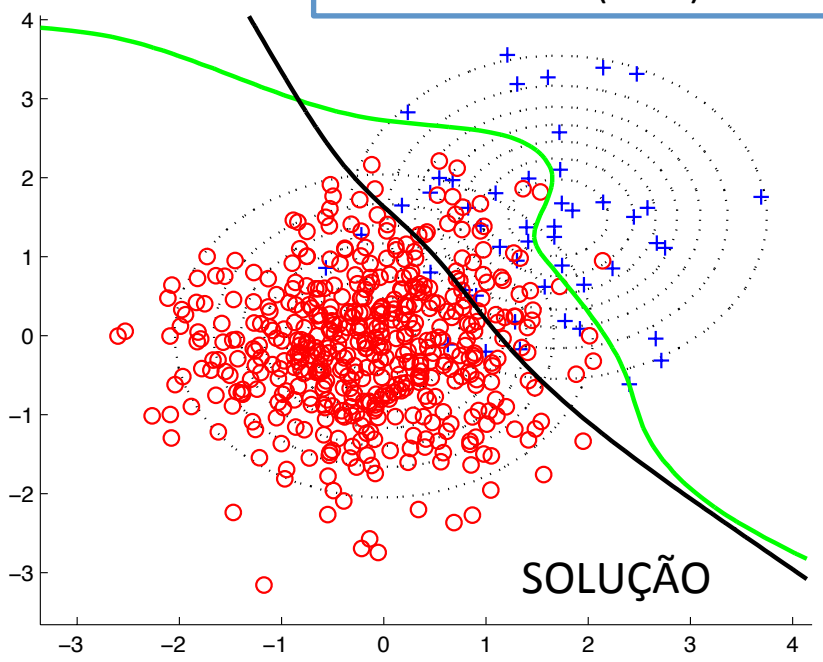
Consequência: baixa taxa de reconhecimento (acurácia) para a classe minoritária que corresponde ao grupo de interesse na maior parte das aplicações reais.

Problema de Classes Desbalanceadas

Taxa de Verd. Positivos: 60%
Taxa de Verd. Negativos: 99%
Curva ROC[1] (Área): 88%



Taxa de Verd. Positivos: 90%
Taxa de Verd. Negativos: 89%
Curva ROC (Área): 94%



1. *Receiver Operating Characteristic*.

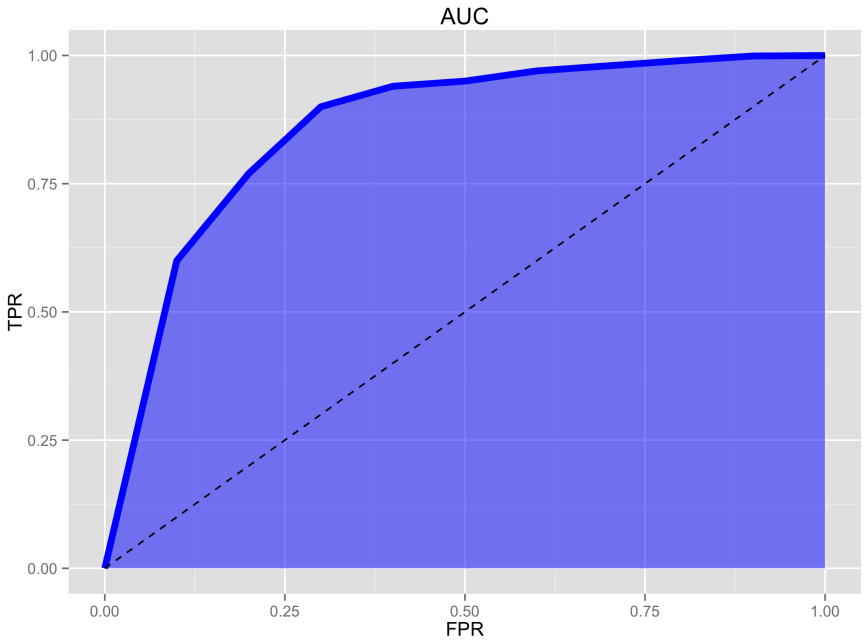
Métricas de Avaliação Mais Comuns

Kubat's G-MEAN:

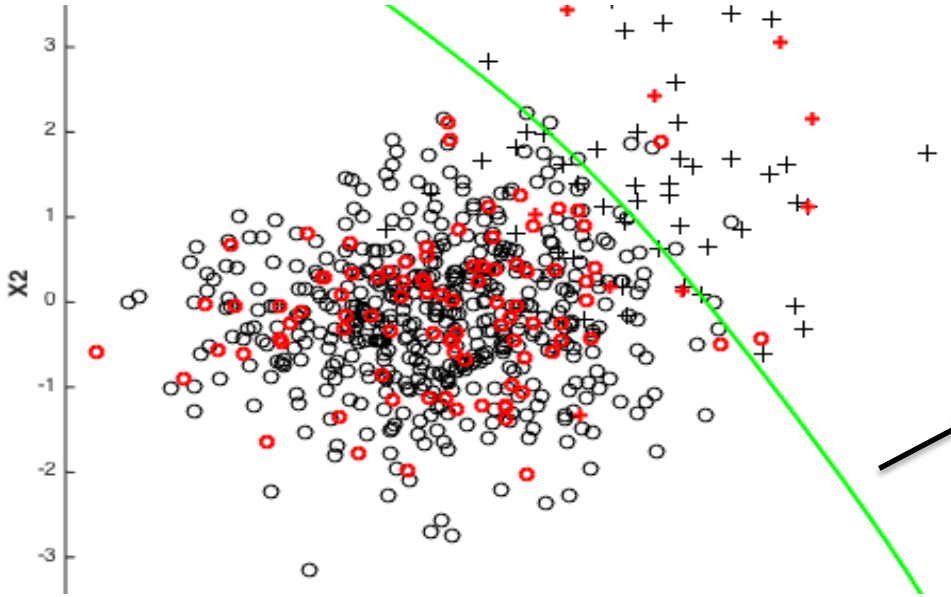
$$GMEAN = \sqrt{TPR \times TNR}$$

TPR = taxa de verdadeiros positivos
TNR = taxa de verdades negativos

AUC (Area Under the ROC Curve)

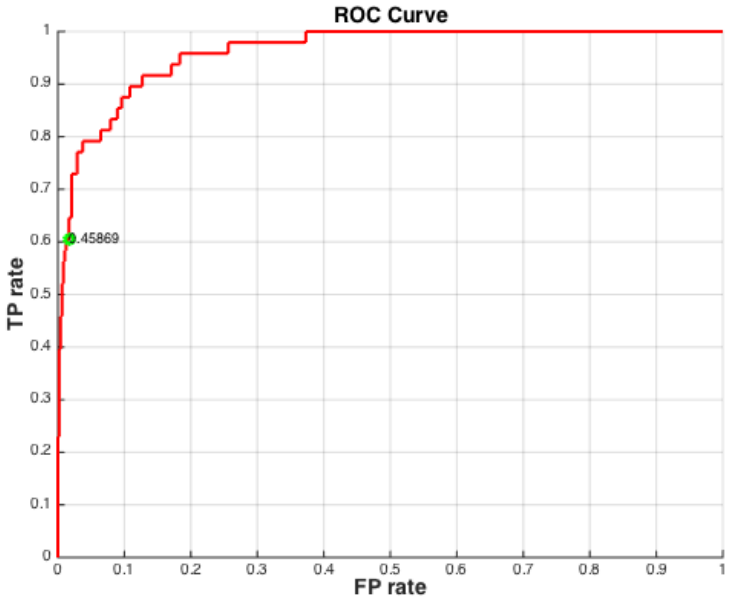


Exemplo – Rede MLP (Topologia 2:3:1)



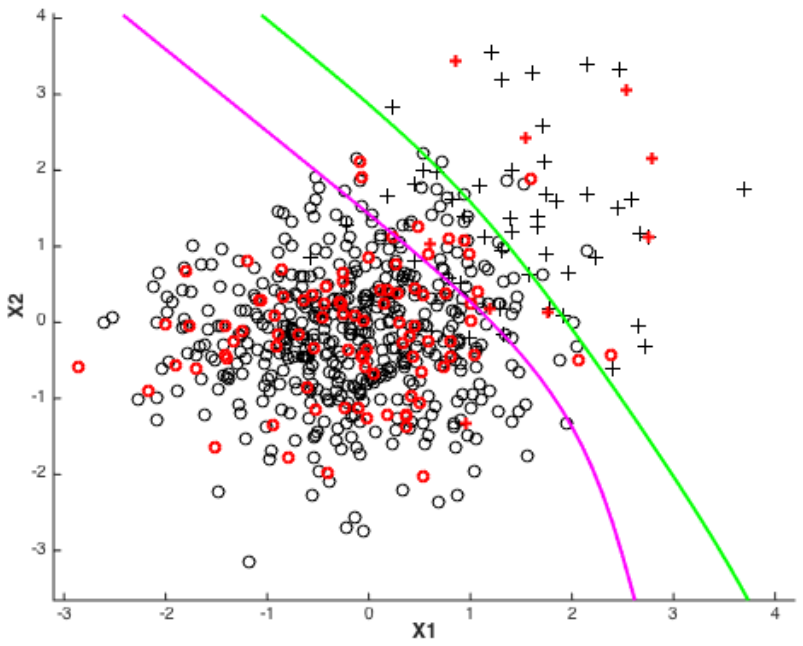
Training set in black
Test set in red

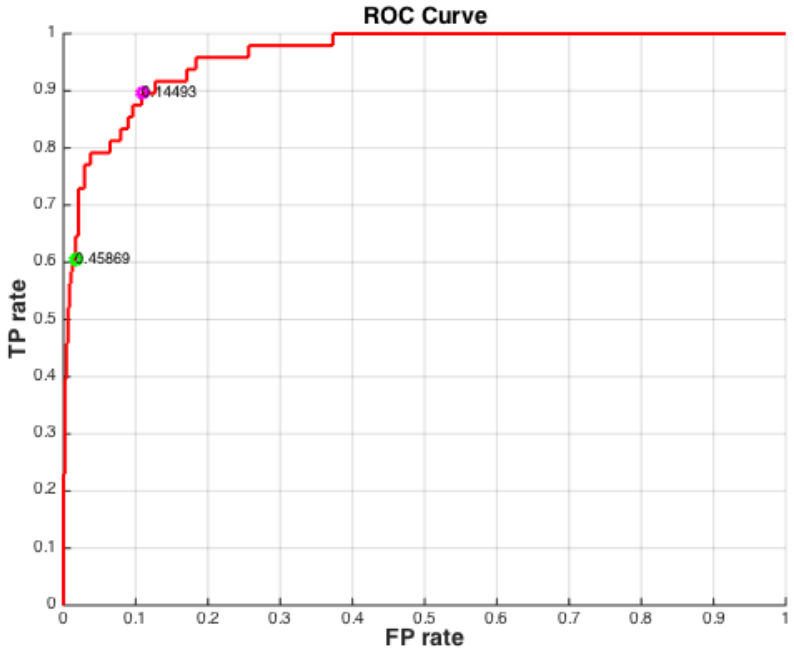
<u>Confusion Matrix (Test Set)</u>
TPR = 0.56 FNR = 0.44
FPR = 0.02 TNR = 0.98



<u>Evaluation Metrics (Test Set)</u>
G-mean: 0.74
AUC: 0.93

Exemplo – Rede MLP (Topologia 2:3:1) – Mudança de Limiar





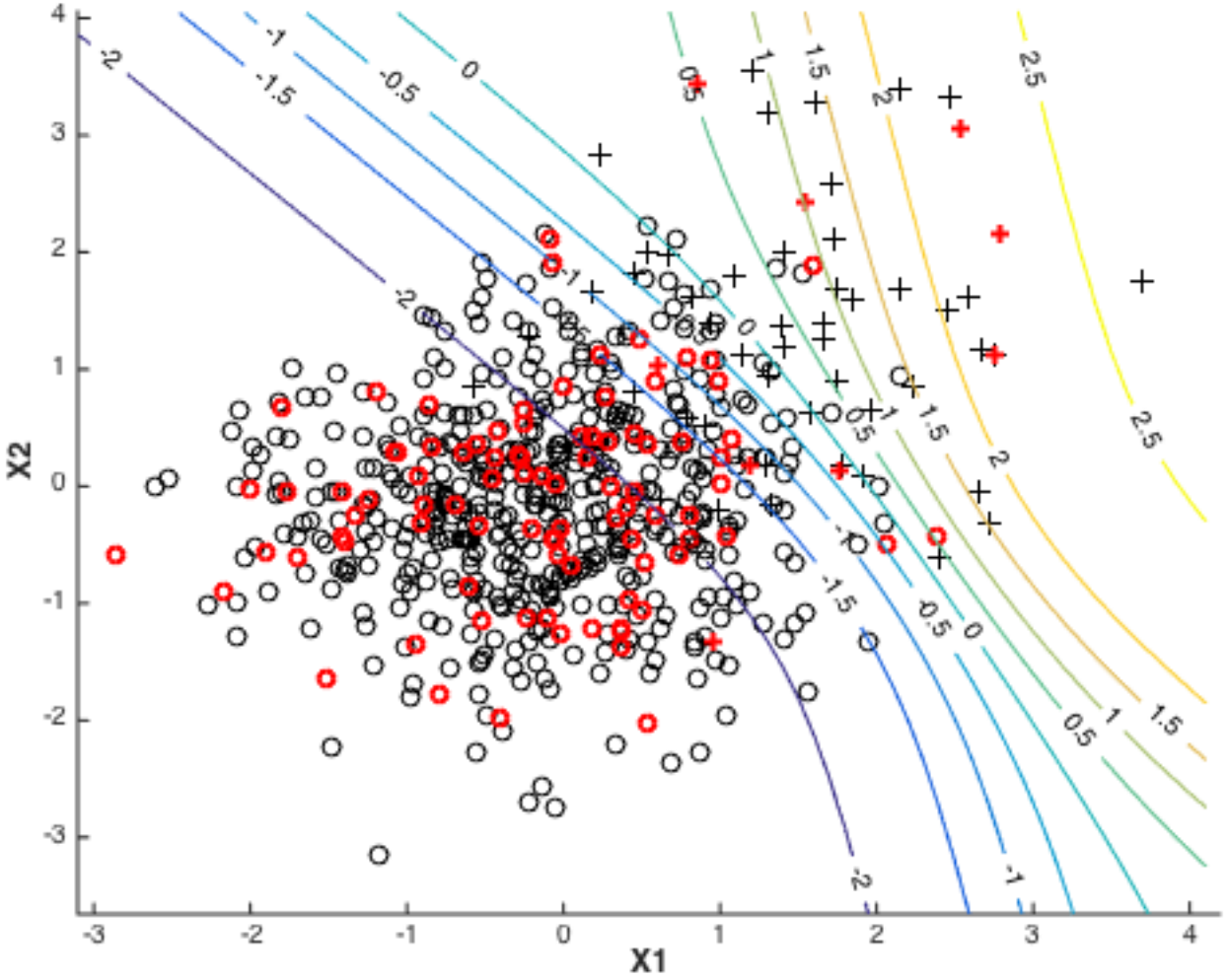
Confusion Matrix (Test Set)
TPR = 0.56 FNR = 0.44
FPR = 0.02 TNR = 0.98

Metrics (Test Set)
G-mean: 0.748
AUC: 0.93

Confusion Matrix (Test Set)
TPR = 0.89 FNR = 0.11
FPR = 0.12 TNR = 0.88

Metrics (Test Set)
G-mean: 0.885
AUC: 0.93

Exemplo - Rede MLP (Topologia 2:3:1): Limiares



O Que Fazer?

- a idéia geral do TP é propor e avaliar uma estratégia para lidar com o Problema de Classes Desbalanceadas.
  - usar rede MLP com classificador base;
  - melhorar o desempenho em relação as métricas:
    - **GMEAN**
    - **AUC**

Guidelines

- **Passo 1**: criar um *baseline* a partir de um conjunto de bases de dados desbalanceadas:

Dataset	alias	# attributes	N_1	N_2	$N_1/(N_1+N_2)$
Ionosphere	iono	34	126	225	0.359
Pima Indians Diabetes	pid	08	268	500	0.349
German Credit	gmn	24	300	700	0.300
WP Breast Cancer	wpbc	33	47	151	0.237
Vehicle (4 vs. all)	veh	18	199	647	0.235
SPECTF Heart	hrt	44	55	212	0.206
Segmentation (1 vs. all)	seg	19	30	180	0.143
Glass (7 vs. all)	gls7	10	29	185	0.136
Euthyroid (1 vs. all)	euth	24	238	1762	0.119
Satimage (4 vs. all)	sat	36	626	5809	0.097
Vowel (1 vs. all)	vow	10	90	900	0.091
Abalone (18 vs. 9)	a18-9	08	42	689	0.057
Glass (6 vs. all)	gls6	10	9	205	0.042
Yeast (9 vs. 1)	y9-1	08	20	463	0.041
Car (3 vs. all)	car	06	69	1659	0.040
Yeast (5 vs. all)	y5	08	51	1433	0.034
Abalone (19 vs. all)	a19	08	32	4145	0.008

Exemplo: Bases de Dados Desbalanceadas

Guidelines

- **Passo 1**: criar um *baseline* a partir de um conjunto de bases de dados desbalanceadas:

Baseline: Redes MLP sem qualquer modificação

Dataset	MLP	SMTTL	WWE	ST2	RBoost	CSMLP
iono	84.33	85.20	85.21	84.68	**86.39**	85.60
pid	69.18	66.81	72.61	70.45	72.12	**72.82**
gmn	57.60	61.05	67.42	**70.08**	65.86	70.00
wpbc	62.07	61.33	61.72	62.80	**66.63**	63.47
veh	95.41	95.91	96.00	96.75	**97.07**	96.84

Exemplo*: Valores Médios de G-mean

Fonte: Castro and Braga. A Novel Cost-Sensitive Approach to Improve the MLPs Performance on Imbalanced Data. IEEE Transactions on Neural Networks and Learning Systems, 2013.

Guidelines

- **Passo 2**:
  - Buscar artigos na literatura que propuseram soluções promissoras para o problema;
    - usar artigos recentes e de “boas fontes” (periódicos e conferências reconhecidas na área de Machine Learning);
  - Implementar/testar soluções;
  - modificar/propor solução;
  - avaliar sua solução em relação ao **baseline** (redes MLP sem qualquer modificação);

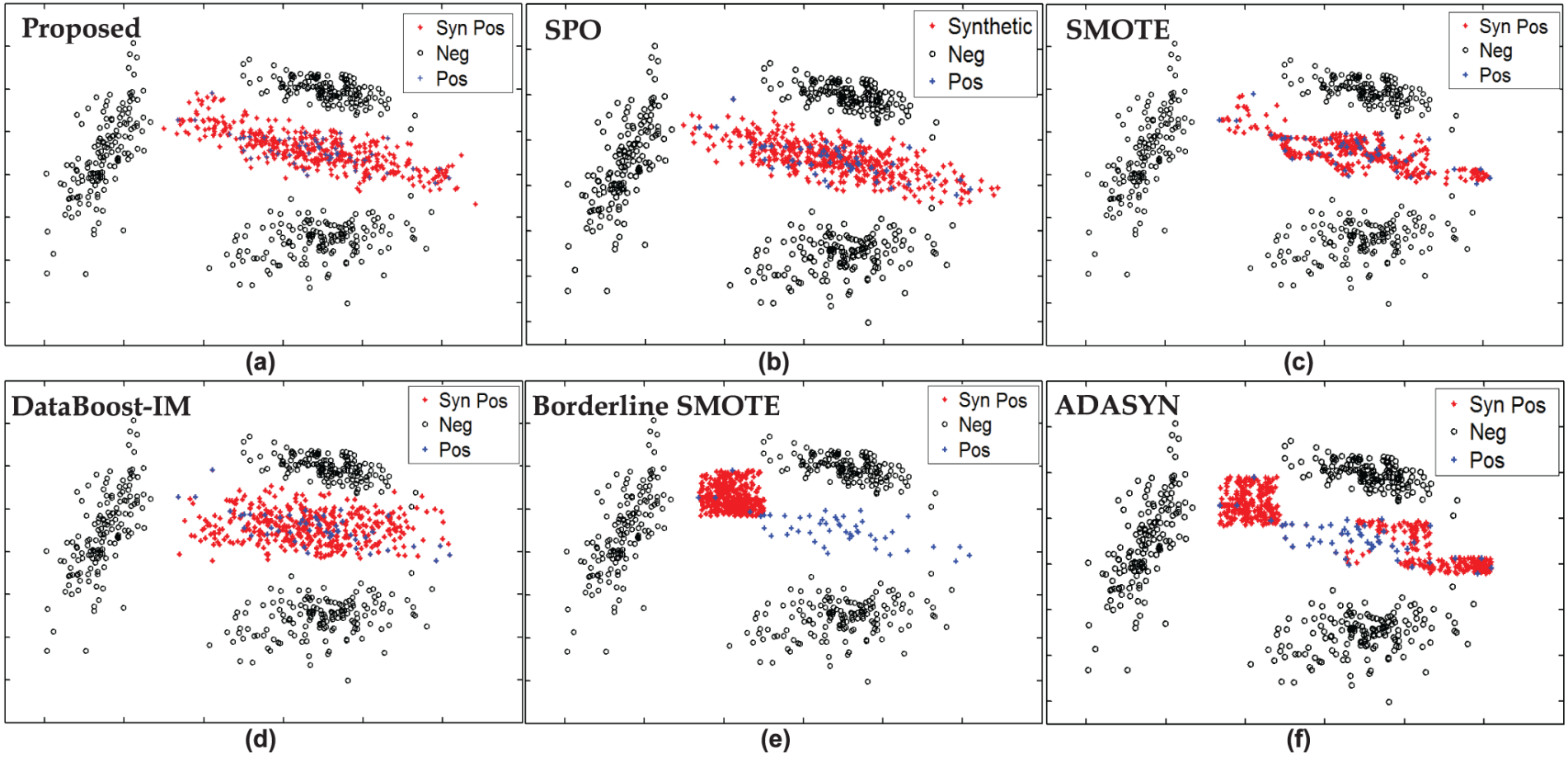
Guidelines

- **Passo 3**: Escrita do Artigo

- Sugestão de Estrutura:

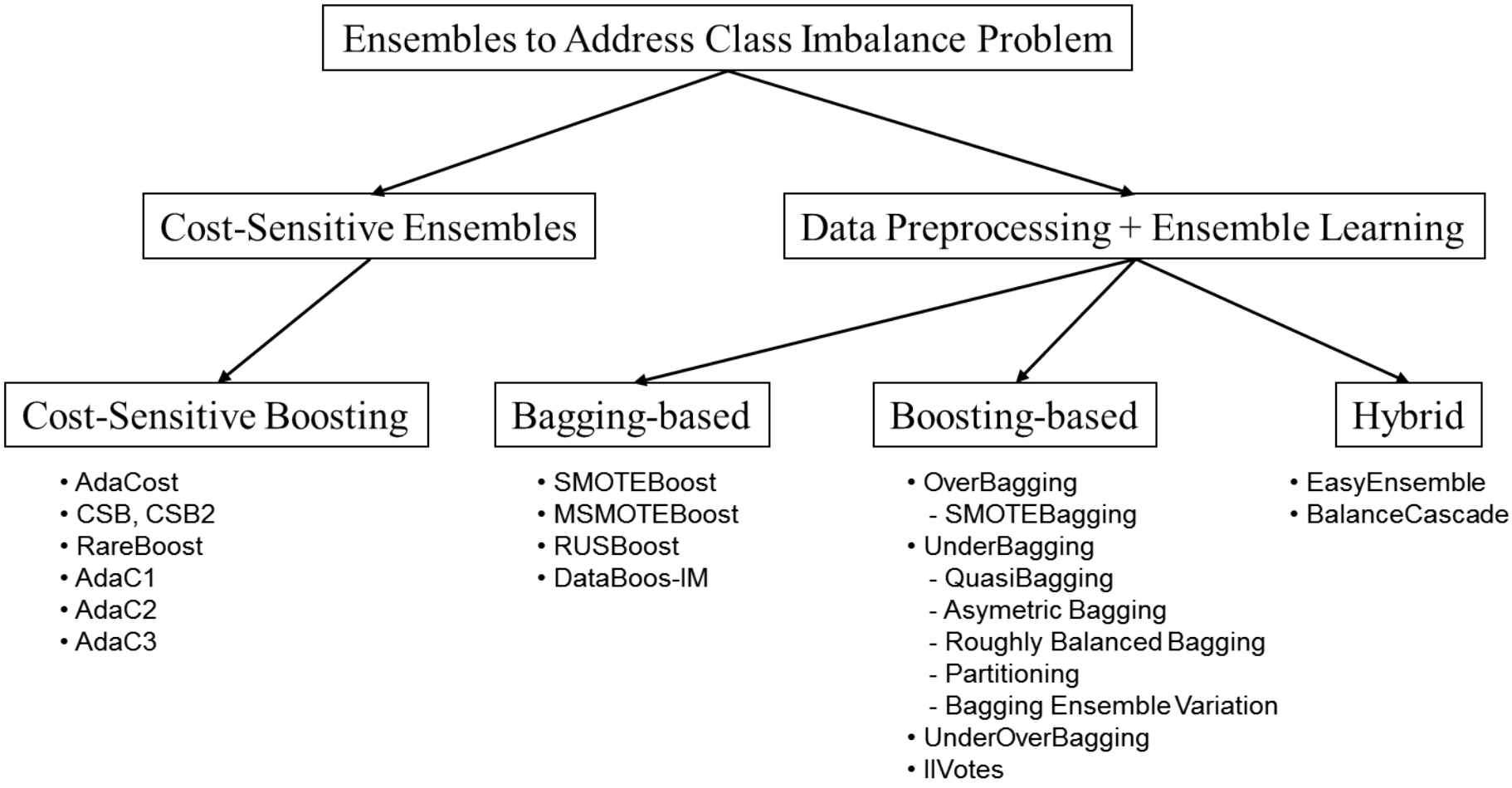
  - **Introdução**
    - contextualização do problema; como o problema tem sido abordado na literatura; abordagem/ideias propostas para solucionar o problema;
  - **Materiais e Métodos:**
    - descrição da abordagem proposta e/ou das abordagens que fundamentam a proposta; metodologia experimental; metricas de avaliação;
  - **Resultados e Discussão**
    - tabelas e gráficos comparativos; discussão dos resultados (era o esperado? sim, não e por que.)
  - **Conclusão**

Exemplos de Soluções - Oversampling



Fonte: Cao et al. Integrated Oversampling for Imbalanced Time Series Classification . IEEE Transactions on Knowledge and Data Engineering, 2013.

Exemplos de Soluções (2) - Ensembles



Temas Sugeridos (29/09)

- RAMOBoost: Ranked Minority Oversampling in Boosting [IEEE-TNNLS, 2010]
- Imbalanced evolving self-organizing learning [Neurocomputing – 2014]
- Cost-sensitive boosting for classification of imbalanced data [Pattern Recognition, 2007]
- PCA and Gaussian noise in MLP neural network training improve generalization in problems with small and unbalanced data sets [IJCNN, 2011]
- Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification [IEEETNN, 2013]
- Boosted SVM with active learning strategy for imbalanced data [Soft Computing, 2014]