Soluciones consistentes

para

problemas de clasificación desequilibrada

Aníbal R. Figueiras-Vidal
CU, UC3M
AN, RAIng

Contenido

1. Clasificación desequilibrada
2. Problemas reales
3. Análisis (Bayes) y máquinas
4. Reequilibrado
5. Equivalencia LR
6. Divergencias de Bregman
7. Reequilibrado neutral
8. Ejemplos
9. Reequilibrado en dos pasos
10. Ejemplos
11. Otro reequilibrado en dos pasos
12. Ejemplos
13. Conclusiones
14. Extensiones

1. Clasificación desequilibrada

❖ Clasificación:

$$\boldsymbol{x} \rightarrow C_j$$

(análogos: - Tests de hipótesis

- Toma de decisiones

…)

❖ Desequilibrio (IB):

Las poblaciones de las clases o/y los costes de clasificación son muy distintos

2. Problemas reales

- Seguridad: Intrusión, alarmas, reconocimiento/identificación…
- Finanzas: Fraude, riesgo, crédito…
- Empresa: Pérdida de clientes, marketing…
- Industria: Averías, defectos, mantenimiento…
- Salud: Diagnóstico, personalización…
- Espacio: Exploración, previsión…
- Economía: Tendencias, desviaciones, anomalías…
- DSP: Textos, lenguaje natural, imagen/vídeo (e.g., objetos)…
- “Smart Soc”: Emergencias, anticipación, servicios…
- Redes sociales: (como arriba)

… y más (Gobierno, Educación…)

3. Análisis (Bayes) y máquinas

(Binario)

❖ Regla de Bayes

$$q_L(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_0)} \underset{C_0}{\overset{C_1}{\gtrless}} \frac{c_{10}-c_{00}}{c_{01}-c_{11}} \frac{P_0}{P_1} = Q_C Q_P = Q \qquad (Q \in (\infty, 0]: \quad \text{NP OG})$$

IB: $Q \gg 1$ (Q_C y/o $Q_P \gg 1$)

❖ Máquina: $\{\boldsymbol{x}^{(n)}, t^{(n)}\}$

- Generativas: $\hat{p}(\boldsymbol{x}|C_i)$

  insensibles / <u>bajas prestaciones</u>

- Discriminativas

  $o_{\boldsymbol{w}}(\boldsymbol{x}) \gtrless \eta$, vía un coste subrogado: $\boldsymbol{w} = \arg\min_{\boldsymbol{w}'} \sum_n c\left(t^{(n)}, o_{\boldsymbol{w}'}\left(\boldsymbol{x}^{(n)}\right)\right)$

  altas prestaciones / <u>sesgados</u> $(\eta \in [1, -1]: \quad \text{MOC})$

4. Reequilibrado

- Modificando los costes
- Re-muestreando (sub-, sobre-) (+ diversidad)
- Generación (+ diversidad)
- Máquinas de una clase
- Aprendizaje activo
- Métricas “ad hoc”

…

Pueden combinarse

… Pero la mayoría de ellos son (puramente) empíricos

5. Equivalencia LR

- Si se consigue (un buen) $\hat{q}_L(\boldsymbol{x})$

  no hay dificultad: $\hat{q}_L(\boldsymbol{x}) \gtrless Q, \quad \forall\, Q$ (estimación Neyman-Pearson OC)

  (→ Es invariante!)

- Se tiene

$$Pr(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)P_1}{p(\boldsymbol{x}|C_1)P_1 + p(\boldsymbol{x}|C_0)P_0} = \frac{q_L(\boldsymbol{x})}{q_L(\boldsymbol{x}) + Q_P}$$

($Q_P \rightarrow Q$: cambiando costes)

$$\Rightarrow q_L(\boldsymbol{x}) = Q\,\frac{Pr(C_1|\boldsymbol{x})}{1 - Pr(C_1|\boldsymbol{x})} \qquad \text{1:1}$$

- Luego basta con $\widehat{Pr}(C_1|\boldsymbol{x})$!

6. Divergencias Bregman (1)

Si el coste surrogado *c* en el entrenamiento es una d.B., i.e.,

si y sólo si $\frac{\partial \mathrm{c}(t,o)}{\partial o} = -g(o)(t-o), \qquad g(\cdot) > 0$

$\Rightarrow \underline{o(\boldsymbol{x}) = \tilde{E}(t|\boldsymbol{x})}$

Con $t = \pm 1$

$E(t|\boldsymbol{x}) = 1\,Pr(C_1|\boldsymbol{x}) - 1\,Pr(C_0|\boldsymbol{x}) = 2\,Pr(C_1|\boldsymbol{x}) - 1$

$\Rightarrow Pr(C_1|\boldsymbol{x}) = \frac{1+o(\boldsymbol{x})}{2}$

Pero $Pr(C_1|\boldsymbol{x})\downarrow\downarrow$ en casos desequilibrados

⇒ Reequilibrado **neutral**

6. Divergencias Bregman (2)

❖ Demostración:

$$\bar{c}(o) = \int c(t,o)p(t|\boldsymbol{x})\,dt$$

$$\frac{\partial \bar{c}(t,o)}{\partial o} = -g(o)\int (t-o)p(t|\boldsymbol{x})\,dt = 0$$

$$\int o\; p(t|\boldsymbol{x})\,dt = o = \int t\; p(t|\boldsymbol{x})\,dt = E(t|\boldsymbol{x})$$

(Q_C puede incluirse con Q_P mediante

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}'}\left[\sum_{n_0} Q_c\, c\left(t^{(n_0)}, o^{(n_0)}\right) + \sum_{n_1} c\left(t^{(n_1)}, o^{(n_1)}\right)\right] \quad)$$

7. Reequilibrado neutral

Se pueden utilizar:

- costes
- remuestreo neutral (sobre- o/y sub-), y conjuntos (~ “bagging”)
- generación neutral (como Parzen, SMOTE…) y conjuntos

y sus combinaciones

No: re-muestreo o generación informados (más en una región determinada)

Entonces:

$$\widetilde{Pr}(C_1|\boldsymbol{x}) = \frac{1+o(\boldsymbol{x})}{2}$$

$$\Rightarrow \tilde{q}_L(\boldsymbol{x}) = \tilde{Q}\,\frac{1+o(\boldsymbol{x})}{1-o(\boldsymbol{x})} = \hat{q}_L(\boldsymbol{x})\ ! \qquad \tilde{Q}\,\frac{1+o(\boldsymbol{x})}{1-o(\boldsymbol{x})} \gtrless_{C_0}^{C_1} \hat{Q}$$

$$\Rightarrow \quad o(\boldsymbol{x}) \gtrless_{C_0}^{C_1} \frac{\hat{Q}-\tilde{Q}}{\hat{Q}+\tilde{Q}} \qquad \left(\text{y} \quad \widehat{Pr}(C_1|\boldsymbol{x}) = 1/\left[1+\frac{\hat{Q}_P}{\tilde{Q}}\,\frac{1-o(\boldsymbol{x})}{1+o(\boldsymbol{x})}\right] \right)$$

8. Ejemplos (1)

❖ No neutral

Problema (sintético)

$$p(\boldsymbol{x}|C_1) = \begin{cases} (1+x_2)/4 & x_1, x_2 \in [-1,1] \\ 0 & \text{e o } c \end{cases}$$

$$p(\boldsymbol{x}|C_0) = \begin{cases} 1/3.6 & x_1 \in [-0.9, 0.9], x_2 \in [-1,1] \\ 0 & \text{e o } c \end{cases}$$

$\Rightarrow P_D = 0.775 - 0.45\,(1 - 2P_{FA}) - 0.225\,(1 - 2P_{FA})^2$

$N_1 = 300, N_0 = 19{,}200 \;\; (IR = N_0/N_1 = 64)$

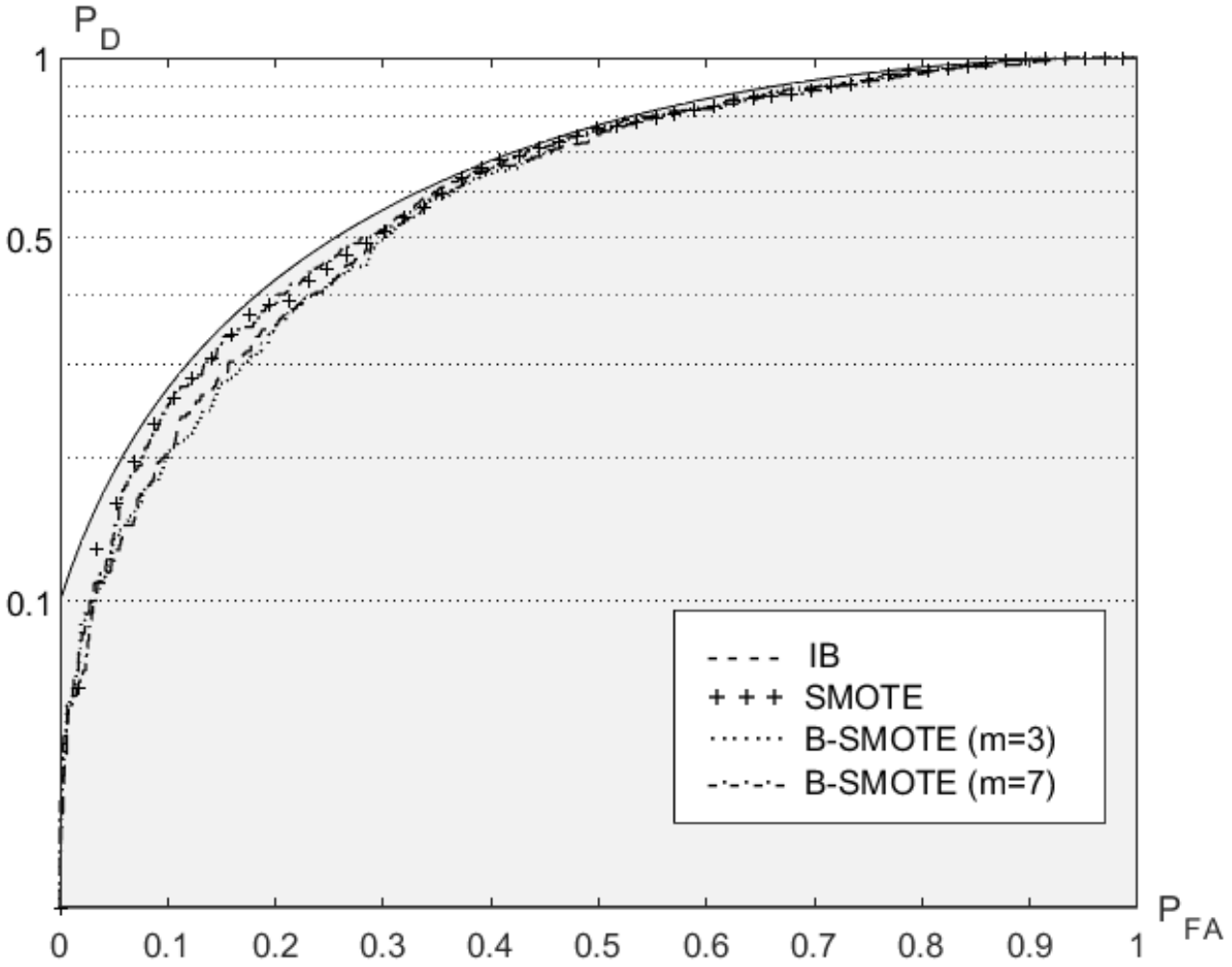
Se prueban, con $c(t,o) = (t-o)^2$:

- Directo (IB)
- SMOTE ($K = 3$): Generación neutral hacia los K NN
- Borderline-SMOTE ($K = 3$): Generación informada: sólo para las muestras con al menos $m/2$ de sus NN mayoritarias (y no m)

(Reequilibrado completo, conjunto de 11 MLPs $H = 4$)

8. Ejemplos (2)

Resultados:



8. Ejemplos (3)

❖ No Bregman

(Problema: “Electricity” $N_0 = 26075$

$N_1 = 19237$ $D = 8$

todas para estimar NP-ROC con 80/20)

Reducido y desequilibrado: $N_0 = 1629$

$N_1 = 33$ $(IR = 50)$

Estimando NP-ROC con un complemento

$c(t, o) = \alpha(t - o)^2 + (1 - \alpha)|t - o|$

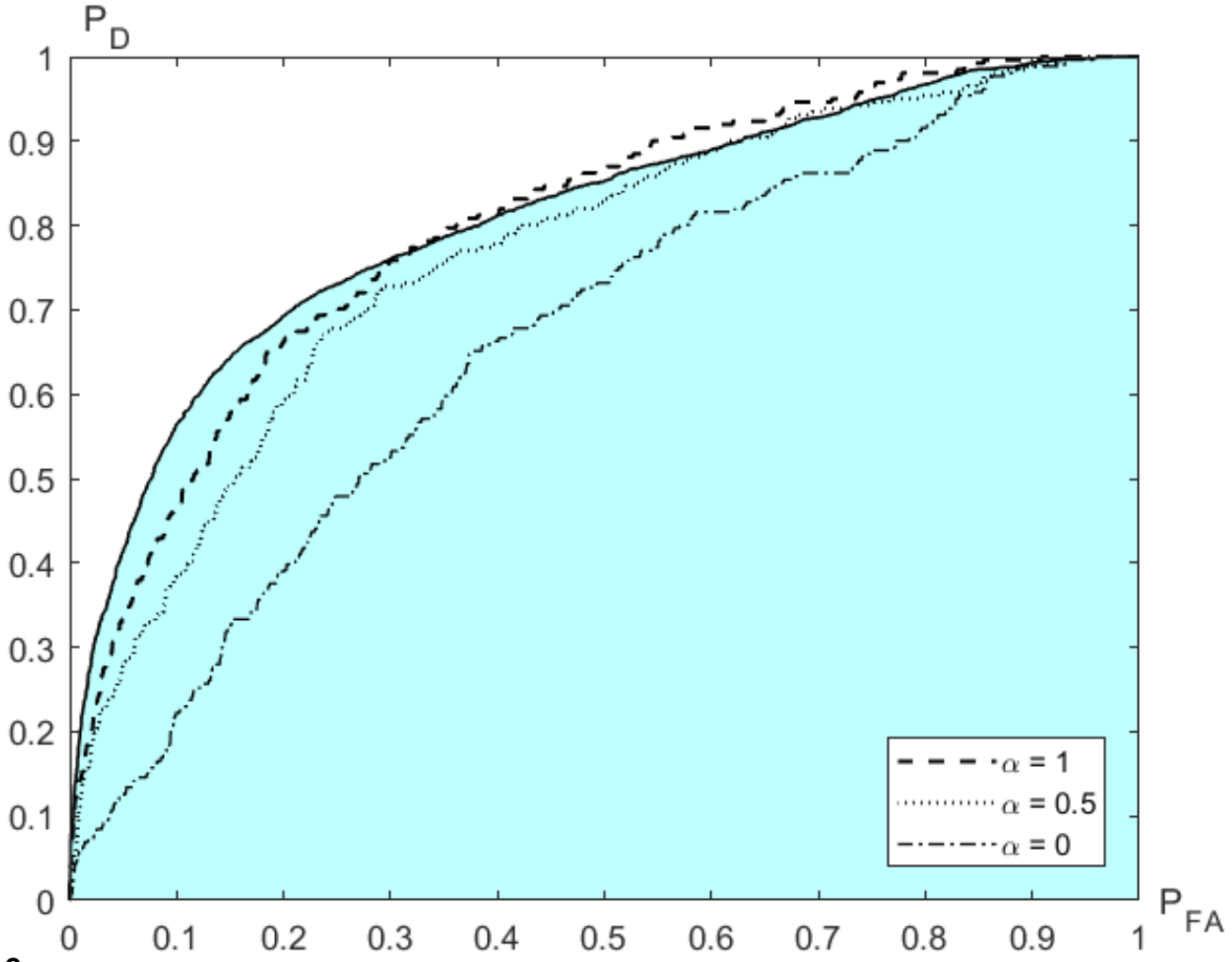
(Bregman con $\alpha = 1$)

y SMOTE para reequilibrado (completo)

(MLP $H = 7$)

8. Ejemplos (4)

Resultados:



9. Reequilibrado en dos pasos (1)

Si –con divergencias de Bregman– se enfatiza según $e_i(\boldsymbol{x})$

se trabaja con

$$q_L'(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)e_1(\boldsymbol{x})/a_1}{p(\boldsymbol{x}|C_0)e_0(\boldsymbol{x})/a_0} = q_L(\boldsymbol{x})q_E(\boldsymbol{x})/Q_A$$

donde a_i son constantes de normalización

$$q_E(\boldsymbol{x}) = e_1(\boldsymbol{x})/e_0(\boldsymbol{x})$$
$$Q_A = a_1/a_0$$

Luego, $q_L(\boldsymbol{x}) = Q_A\, q_L'(\boldsymbol{x})/q_E(\boldsymbol{x})$

9. Reequilibrado en dos pasos (2)

Trabajando con $\hat{q}'_L(\boldsymbol{x})$, se obtendrá $\hat{q}'_L(\boldsymbol{x}) = \tilde{Q}_2(\boldsymbol{x}) \dfrac{1+o_{B2}(\boldsymbol{x})}{1-o_{B2}(\boldsymbol{x})}$

de donde el test

$$\frac{\tilde{Q}_2(\boldsymbol{x})}{q_E(\boldsymbol{x})} \frac{1+o_{B2}(\boldsymbol{x})}{1-o_{B2}(\boldsymbol{x})} \underset{C_0}{\overset{C_1}{\gtrless}} \frac{\hat{Q}}{\hat{Q}_A}$$

Si –como es habitual– se entrena el segundo clasificador bajo un criterio MAP (mínima probabilidad de error) una vez que se ponderan las muestras, resulta obvio que $\tilde{Q}_2(\boldsymbol{x}) = q_E(\boldsymbol{x})$. Por ello, el test resulta

$$\frac{1 + o_{B2}(\boldsymbol{x})}{1 - o_{B2}(\boldsymbol{x})} \underset{C_0}{\overset{C_1}{\gtrless}} \frac{\hat{Q}}{\hat{Q}_A}$$

i.e.

$$o_{B2}(\boldsymbol{x}) \underset{C_0}{\overset{C_1}{\gtrless}} \frac{\hat{Q} - \hat{Q}_A}{\hat{Q} + \hat{Q}_A} = \eta_2$$

9. Reequilibrado en dos pasos (3)

Por otro lado, pueden estimarse las a_i con las formas muestrales

$$\hat{a}_i = \frac{1}{N_i} \sum_{n_i} e_i\left(\boldsymbol{x}^{(n_i)}\right)$$

Estos estimadores permiten aplicar el test para una política de costes dada.

Remuestreo o/y generación no toleran pesos arbitrarios, y se debe aplicar diversidad (lo cual es una ventaja) y promediado.

Cuando se utiliza generación

$$\hat{a}_i = \frac{1}{N_i'} \sum_{n_i'} e_i'\left(\boldsymbol{x}^{(n_i')}\right)$$

donde $e_i'(\cdot)$ excluye la tasa de generación, $\left\{\boldsymbol{x}^{(n_i')}\right\}$ son cada muestra $\boldsymbol{x}^{(n_i)}$ y las generadas a partir de ella, y N_i' el número total de ejemplos.

La forma anterior es aceptable si las muestras generadas están cerca de las originales.

9. Reequilibrado en dos pasos (4)

El proceso informado fundamentado (para costes indeterminados) es:

Paso 1:

- Aplicar un reequilibrado neutral y un coste de Bregman para estimar la NP-ROC
- Seleccionar el punto de trabajo (P_{FAW})
- Determinar la ponderación directa y/o tasas de remuestreo/ generación del segundo paso para las muestras según la estimación del NP-ROC y el punto de trabajo

Paso 2:

- Aplicar $e_i(\boldsymbol{x})$ y resolver el segundo problema de reequilibrado, obteniendo η_2 para conseguir P_{FAW}
- Se puede obtener una estimación de la ROC de entrenamiento a partir de ordenar $o_{B2}(\boldsymbol{x})$ con un umbral que vaya desde 1 hasta -1

Por último

- Se pueden clasificar las muestras de test/operación siguiendo la formulación y los mecanismos anteriores

10. Ejemplos (1)

"Bloques" 1 y 5 de la base de datos "BNG: Page-blocks"

(1) C_0: $N_0 = 265{,}174$

(5) C_1: $N_1 = \quad 6{,}238 \qquad\qquad D = 10$

Seleccionamos aleatoriamente 1/3 para el experimento

3/1 entrenamiento : test

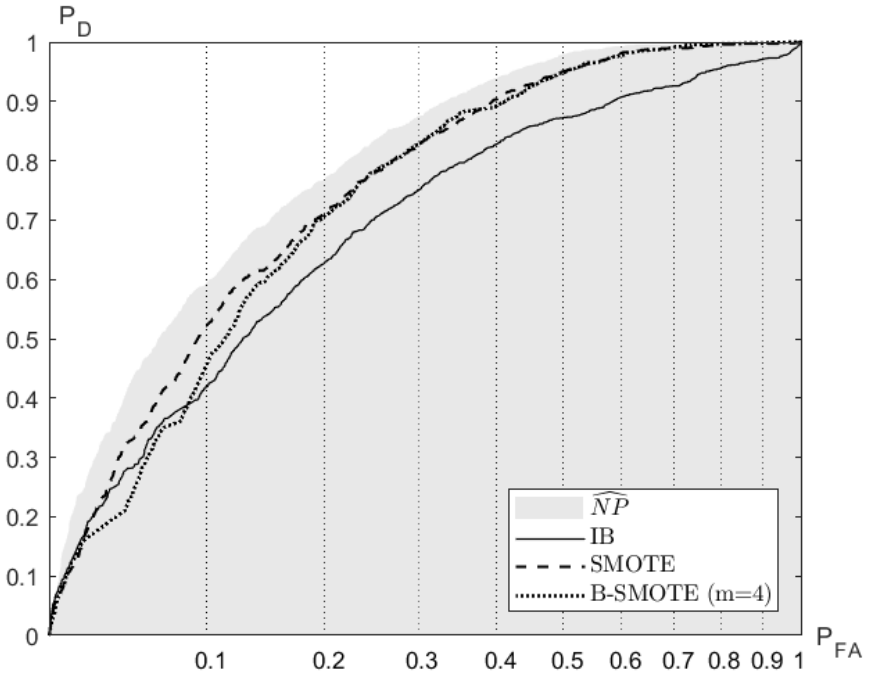
Se utilizan conjuntos de $M = 5$ MLPs (10-27-1)

Se emplea SMOTE ($K = 3$)

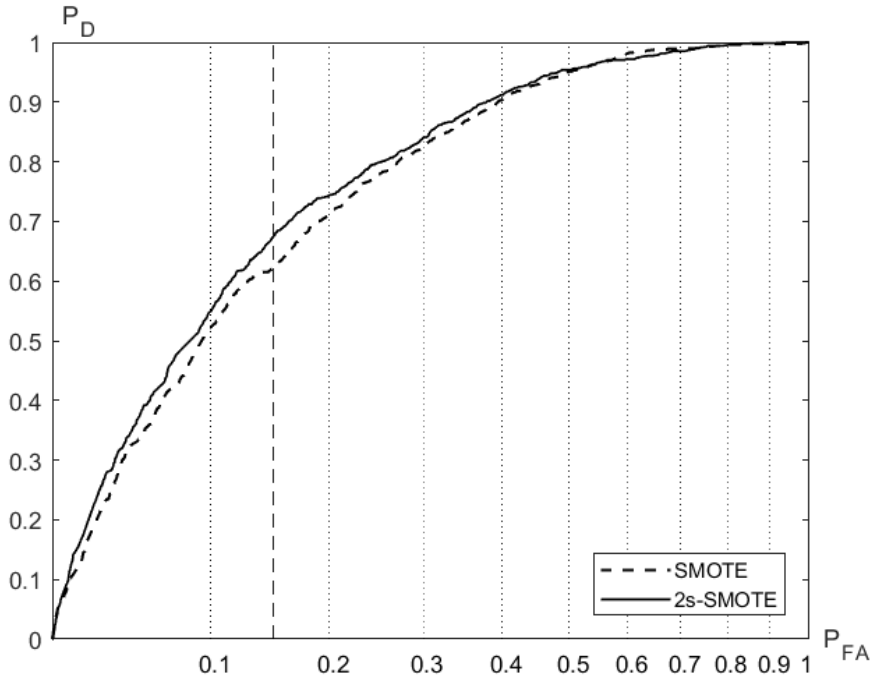
B-SMOTE ($K = 3, m = 4$)

- reequilibrado completo en el primer paso
- reequilibrado completo entre dos umbrales alrededor de P_{FAW} en el segundo paso

10. Ejemplos (2)



Primer paso

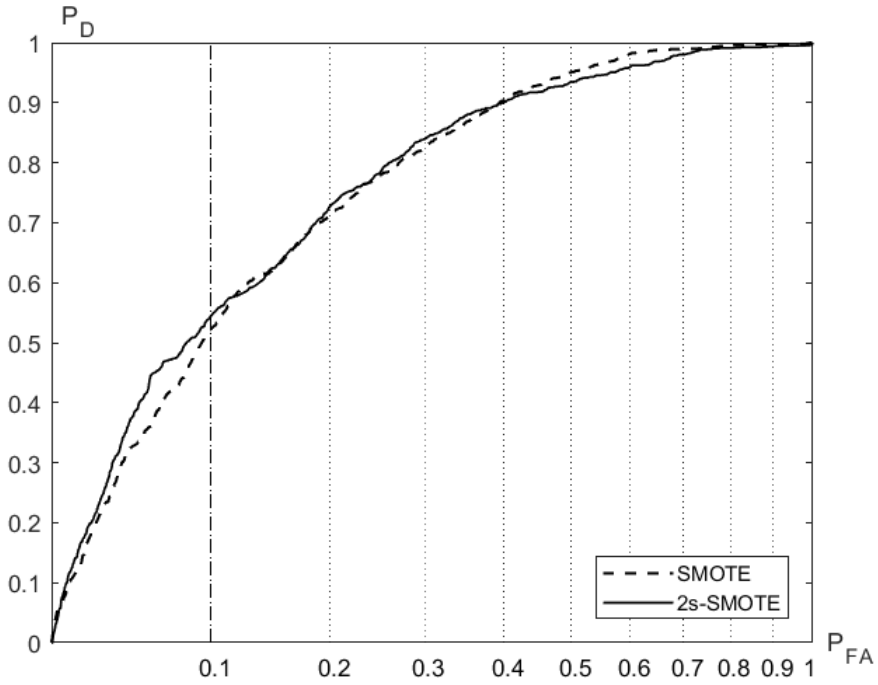


Segundo paso

$P_{FAW} = 0.15$ y umbrales para $P_{FA} = 0$ y $P_{FA} = 0.3$

(Umbrales y énfasis requieren mucho cuidado)

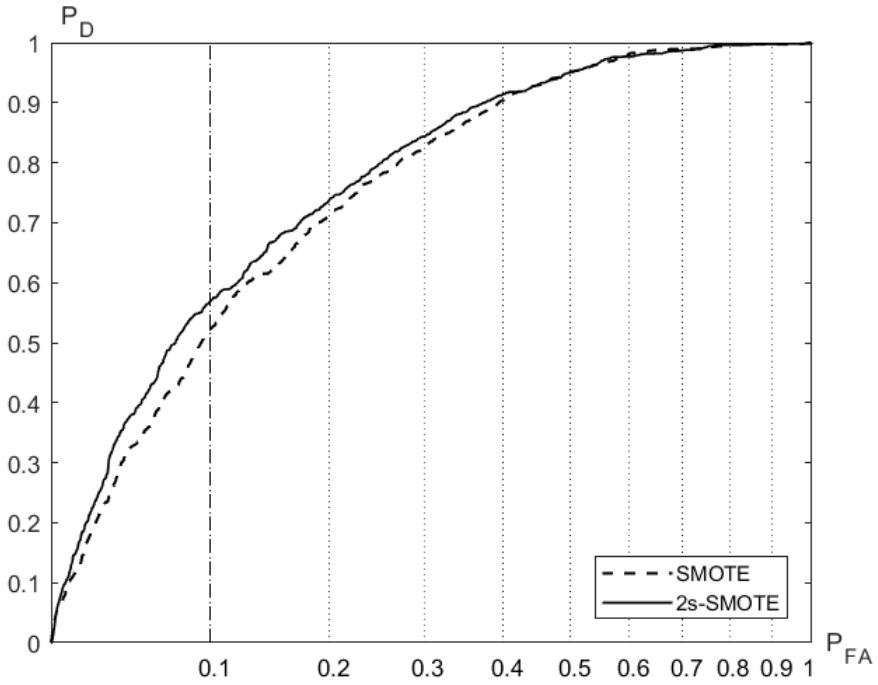
10. Ejemplos (3)



Segundo paso

$P_{FAW} = 0.1$ y umbrales para $P_{FA} = 0$ y $P_{FA} = 0.2$

(Simétrico)



Segundo paso

$P_{FAW} = 0.1$ y umbrales para $P_{FA} = 0$ y $P_{FA} = 0.25$

(Asimétrico)

11. Otro reequilibrado en dos pasos

Versión alternativa: Reequilibrado neutral y enfatizado

Paso 1:

- Mismo proceso que en el método anterior: reequilibrar de forma neutral con costes de Bregman y fijar el punto de trabajo (P_{FAW})
- Establecer una ponderación de la forma:

$$\begin{cases} \hat{Q}_A > 1, & \text{para muestras dentro de un margen en torno a } P_{FAW} \\ 1, & \text{fuera de ese margen} \end{cases}$$

Paso 2:

- (Nuevo) reequilibrado neutral ($\tilde{Q}_2$) y ponderación sobre todas las muestras según $\hat{Q}_A$ fijada en el paso previo

Por último

- Se pueden clasificar las muestras de test/operación siguiendo

$$o_{B2}(\boldsymbol{x}) \underset{C_0}{\overset{C_1}{\gtrless}} \frac{\hat{Q} - \tilde{Q}_2 \hat{Q}_A}{\hat{Q} + \tilde{Q}_2 \hat{Q}_A} = \eta_2'$$

12. Ejemplos (1)

Misma base de datos (“BNG: Page-blocks” 1 vs 5) y entrenamiento / test

Se utilizan conjuntos de $M = 5$ MLPs (10-27-1)

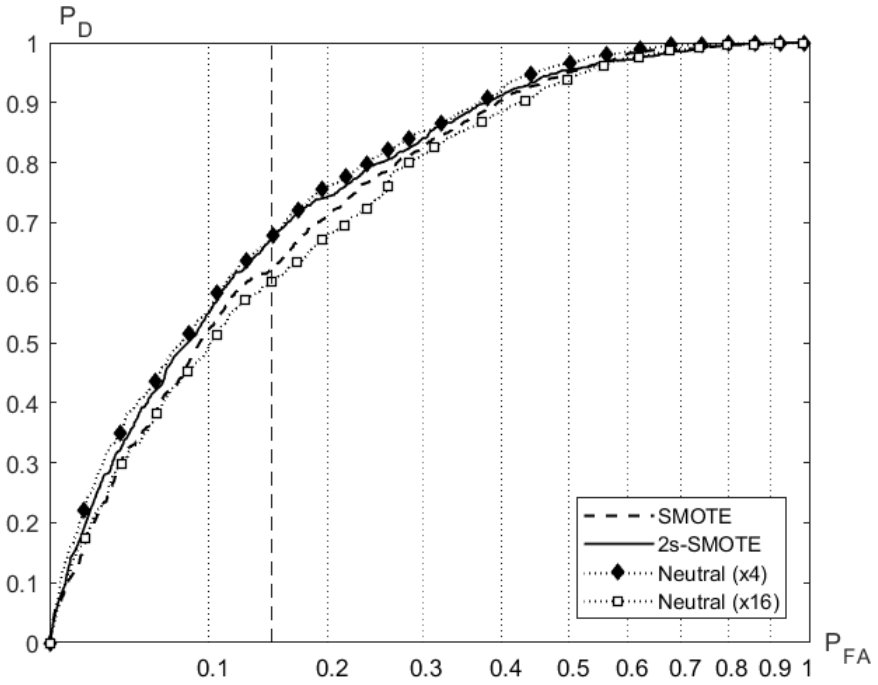
Se emplea SMOTE ($K = 3$)

- reequilibrado completo en el primer paso
- reequilibrado completo en el segundo paso y ponderación de las muestras entre dos umbrales alrededor de P_{FAW}

  (probamos dos niveles de ponderación para analizar sus efectos:

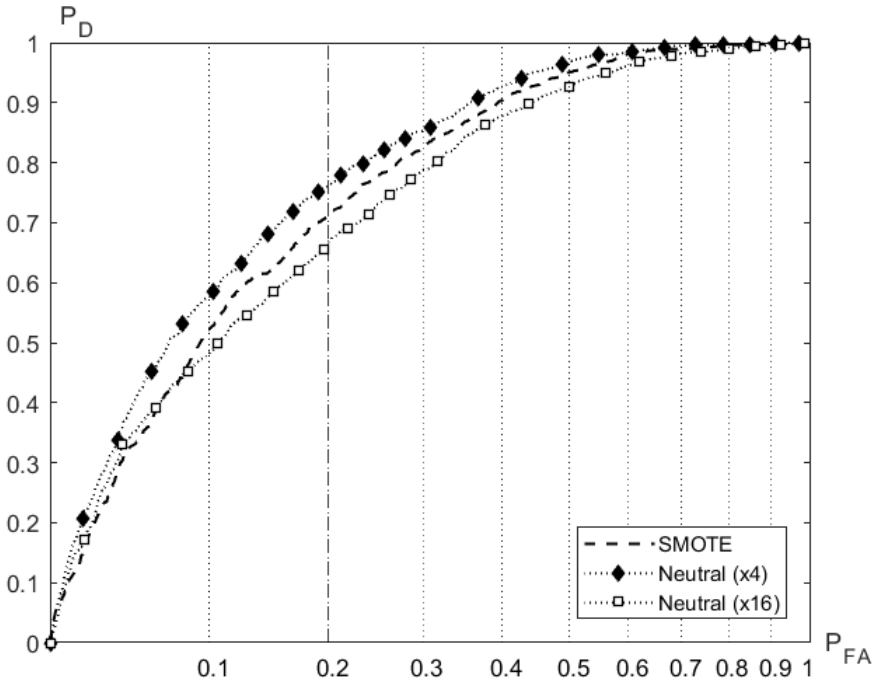
  $\hat{Q}_{A_1} = 4$ y $\hat{Q}_{A_2} = 16$)

12. Ejemplos (2)



Segundo paso

$P_{FAW} = 0.15$ y umbrales para $P_{FA} = 0.1$ y $P_{FA} = 0.2$



Segundo paso

$P_{FAW} = 0.2$ y umbrales para $P_{FA} = 0.15$ y $P_{FA} = 0.25$

13. Conclusiones

- Mediante divergencias de Bregman

  reequilibrados neutrales

  se obtiene una estimación consistente de la ROC NP

  y se puede elegir el punto de trabajo

- Si se desea mejorar la estimación en torno al punto de trabajo

  se determina en un primer paso

  se utiliza un reequilibrado informado en un segundo paso

  (y se compensa)

14. Extensiones

- Extensiones a:
  - Problemas multiclase (directamente o usando dicotomías)
  - Costes funcionales
  - Conjuntos

  (en progreso)
  - Problemas relacionados (ordinal, multi-tarea/multi-etiqueta…) (parcialmente implementado)
- Además,
  - Usando Máquinas Profundas (y “Big Data”) (en progreso)
  - Versiones on-line
  - Añadiendo mecanismos generativos (en progreso)

y aplicaciones del mundo real, por supuesto!

Bibliografía de carácter tutorial

* H. He, E. A. García, “Learning from imbalanced data,” IEEE Trans. Knowledge and Data Eng., vol. 21, pp. 1263–1284, 2009.

* Y. Sun, A.K.C. Wong, M.S. Kamel, “Classification of imbalanced data: A review”, Intl. J. Patter Recognition and Artificial Intell., vol. 23, pp. 687-719, 2009.

* V. López, A. Fernández, 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, vol. 250, pp. 113–141, 2013.

Bibliografía de carácter tutorial

* H. He, Y. Ma (eds.), *Imbalanced Learning: Foundations, Algorithms, and Applications.* Hoboken, NJ: IEEE-Wiley, 2013.

* P. Branco, L. Torgo, R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains”, ACM Computer Surveys, vol. 49, pp. 31:1–31:50, 2016.

* S. Wang, X. Xao, “Multiclass imbalance problems: Analysis and potential solutions,” IEEE Trans. Sys., Man, and Cybernetics – Pt. B: Cybernetics, vol. 42, pp. 1119–1130, 2012.