Clasificación con costes funcionales: soluciones fundamentadas

Aníbal R. Figueiras Vidal
AN, RAIng
et al.

.

Contenido

1. Problemas singulares
2. Clasificación con costes dependientes de la muestra
3. Solución en dos pasos
4. Un ejemplo de aplicación
5. Generalizaciones y extensiones
6. Agradecimientos

1. Problemas singulares

En clasificación (toma de decisiones, test de hipótesis…)

- Desequilibrados
- Con costes dependientes de la muestra
- Ordinales
- Multitarea/multietiqueta

…

Son frecuentes e importantes

Las DLMs convencionales dan malos resultados

Las modificaciones no fundamentadas encierran riesgos

2. Clasificación con costes dependientes del ejemplo

En Negocio y Finanzas:

- detección de fraudes
- calificación de créditos
- marketing personalizado
- gestión de la rotación de clientes
- predicción de quiebras

…

También en Salud, Seguridad…

No puede resolverse de modo fundamentado mediante
modificación directa del coste subrogado

(sí con modificación Bayesiana, pero aprendiendo costes;
sí con algoritmos Bayes-Parzen)

3. Solución en dos pasos (1)

(Para binarios)

Según Bayes:

$$j^* = \arg\min_j \left\{ \sum_i c_{ji}(\mathbf{x}) \Pr(C_i|\mathbf{x}) \right\}$$

de modo que basta $\widehat{\Pr}(C_1|\mathbf{x})$.

Para obtenerla: entrenamiento con divergencia de Bregman

$$\frac{\partial c_B(t,o)}{\partial o} = -g(o)(t-o)$$

como coste subrogado.

Con $t = \pm 1$: $\widehat{\Pr}(C_1|\mathbf{x}) = \dfrac{1 + o(\mathbf{x})}{2}$

3. Solución en dos pasos (2)

De ese modo

- Paso 1: $\widehat{\Pr}(C_1|\mathbf{x})$
- Paso 2: $\hat{j}^* = \arg\min_j \left\{ \sum c_{ji}(\mathbf{x})\, \widehat{\Pr}(C_i|\mathbf{x}) \right\}$

(permite evaluar política de costes;
puede necesitar aprendizaje de costes)

Puede haber también desequilibrio muestral

$\Longrightarrow$ Reequilibrado por invarianza del cociente de verosimilitudes:

a) Reequilibrado neutral: $\widetilde{\Pr}(C_1|\mathbf{x})$

b) Recuperación de $\widehat{\Pr}(C_1|\mathbf{x})$ vía $\tilde{r}(\mathbf{x})$

3. Solución en dos pasos (3)

Proceso para b)

$$\widetilde{\Pr}(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)\,\tilde{P}_1}{p(\mathbf{x}|C_1)\tilde{P}_1 + p(\mathbf{x}|C_0)\tilde{P}_0} = \frac{1}{1+\tilde{Q}_P/\tilde{r}(\mathbf{x})} \qquad \left(\tilde{Q}_P = \tilde{P}_0/\tilde{P}_1\right)$$

$$\tilde{r}(\mathbf{x}) = \tilde{Q}_P \frac{\tilde{P}_r(C_1|\mathbf{x})}{1-\tilde{P}_r(C_1|\mathbf{x})} \quad \left(= \tilde{Q}_P \frac{1-o(\mathbf{x})}{1+o(\mathbf{x})}\right) = \hat{r}(\mathbf{x}) \quad !$$

$$\Longrightarrow \widehat{\Pr}(C_1|\mathbf{x}) = \frac{1}{1+\hat{Q}_P/\tilde{r}(\mathbf{x})}$$

(puede trabajarse con $\tilde{r}(\mathbf{x}) \underset{C_0}{\overset{C_1}{\gtrless}} Q_c(\mathbf{x})\hat{Q}_P$)

4. Un ejemplo de aplicación

Problema: HMEQ (créditos vivienda), IB $\simeq$ 4

Ahorro (promedios en 100 particiones de media ± desviación sobre 100 pases)

– Mejor de Bahnsen (Selva Aleat. + Riesgo Min.)	75’2 ± 1’8
– Segundo de Bahnsen (Árbol de Decisión con Costes)	65’7 ± 0
– Fundamentado 1 MLP subóptimo (RB $\simeq$ 4, pond.)	73’0 ± 1’5
– Fundamentado 1 MLP optimizado (RB = 1)	75’3 ± 1’6
– Fundamentado 31 MLPs optimizado (RB = 3, 100% SMOTE)	78’2 ± 0’6

5. Generalidades y extensiones

G: – a problemas multiclase

E: – a ordinales

– a multitarea/multietiqueta

…

Por esas vías, a <u>diseños insesgados</u>

6. Agradecimientos

A mi equipo de trabajo: Dr. Marcelino Lázaro
Dr. Francisco J. Glez. Serrano
Dra. Lorena Álvarez
Dr. Harold Molina
Ing. Alexander Benítez
Ing. Javier Mediavilla
Ing. Aitor Gutiérrez
Ing. Álvaro Callejas

y colaboradores externos: Prof. Monson Hayes (GMU, USA)
Prof. V. John Mathews (OrSU, USA)
Dr. J.L. Sancho (UPCT)

A las entidades financiadoras: FBBVA
Teldat, S.A.

anibalrfv@tsc.uc3m.es