

Guía 3: técnicas de aprendizaje maquina

Agosto 2019

1. Objetivos

Que el alumno sea capaz de:

- Afianzar los conocimientos sobre técnicas de detección de patrones aprendidos en la instancia teórica.
- Comprender el impacto del proceso de validación y de las diferentes medidas de desempeño utilizadas para estudiar un método de clasificación
- Utilizar los métodos estudiados en problemas de clasificación concretos dentro de BCI.

2. Trabajo Coloquial

1. Los algoritmos de aprendizaje maquina aprenden sus modelos predictivos de un conjunto de datos de entrenamiento. Dicho datos de entrenamiento, puede presentar un desbalance de clase, habiendo entonces más observaciones que pertenecen a una clase en particular. En las BCIs, el paradigma de ERP (P300) es bien conocido por ser una problema de clasificación binario pero sumamente desbalanceado. En particular la relación 'ConP300' vs. 'SinP300' es 1:5. Una investigadora está interesada en conocer el impacto de este desbalance para un clasificador en particular cuando su desempeño es evaluado utilizando diferentes métricas. Para ello realiza un experimento utilizando el mismo clasificador entrenados en dos diferentes configuraciones. En la primera no tiene falsos negativos pero sí falsos positivos, mientras que en la segunda el número de falsos negativos es 30. La investigadora evalúa ambos clasificadores con la misma base de datos utilizando una distribución de datos balanceada y la original (desbalanceada). Los resultados arrojados por cada clasificador para datos balanceados (1:1) y desbalanceados (1:5) se muestra en la Figura 1.


```
---
(sin desbalance)
tp  fp  tn  fn    tpr    pre    tnr    acc    F1      G  AUCROC  AUCPR
500  90  410   0  1.0000  0.8475  0.8200  0.9100  0.9174  0.9206  0.9092  0.8369
470   0  500  30  0.9400  1.0000  1.0000  0.9700  0.9691  0.9695  0.9691  0.9072
---
(desbalance 1:5)
tp  fp  tn  fn    tpr    pre    tnr    acc    F1      G  AUCROC  AUCPR
167  90  743   0  1.0000  0.6498  0.8920  0.9100  0.7877  0.8061  0.9457  0.6440
137   0  833  30  0.8204  1.0000  1.0000  0.9700  0.9013  0.9057  0.9073  0.8249
---
```

Figura 1: Indices de clasificación para dos clasificadores iguales pero con diferentes configuraciones evaluados con: arriba) una base de datos balanceada (1:1) y abajo) una base de datos desbalanceada (1:5). TP: verdaderos positivos, FP: falsos positivos, TN: verdaderos negativos, FN: falsos negativos, TPR: true positive rate, PRE: precisión, TNR: true negative rate, F1: F1-score, G: G-measure, AUCROC: área bajo la curva ROC, AUCPR: área bajo la curva PR.

Al analizar estos resultados, el investigador se encuentra un poco confundido y no sabe cuál es la mejor medida para reportar los resultados de clasificación. Ud. qué piensa al respecto? Indique cuales medidas Ud. utilizaría y cuales no consideraría para el análisis desbalanceado. Justifique su respuesta.

3. Trabajo de Laboratorio

3.1. Análisis discriminante lineal

1. **Utilice LDA como un método de transformación lineal.** Para ello use los datos 'Datos_Sujeto1_256' y 'Datos_Sujeto1_32', separe los datos en entrenamiento y testeo (70/30), y  grafique la proyección de los datos en el espacio de LDA. ¿Qué ocurre cuando utiliza los datos de dimensión 256? ¿Es posible separar linealmente los datos?
2. **Evalúe el impacto del parámetro de regularización en el método Shrinkage-LDA (LDA regularizado).** Para ello utilice los datos 'Datos_Sujeto1_32' en una ventana de tiempo de 250-650 ms luego de aplicado el estímulo. Separe los datos en entrenamiento/validación (90/10). Utilice validación cruzada de 5 particiones en los datos de entrenamiento. Balancee los datos (1:1). Considere ocho (8) valores del parámetro de regularización γ ([0; 0,01; 0,05; 0,1; 0,3; 0,5; 0,8; 1]). Para evaluar el desempeño del clasificador utilice el error de clasificación tanto en entrenamiento, testeo y validación. Represente en una misma gráfica el error de entrenamiento, testeo y validación en función del parámetro de regularización. Responda: ¿para qué valor del parámetro de regularización el clasificador obtiene el mejor rendimiento? ¿Cree que este parámetro será el mismo si evalúa el método en otro sujeto diferente?
3. Es muy deseable que el algoritmo de decodificación utilizado en la BCI sea capaz de obtener resultados *confiables* aunque se los entrene con pocos datos de entrenamiento. Esto permitirá que el tiempo de calibración pueda disminuirse, y así aumentar la practicidad y uso de los sistemas BCIs.

Compare el desempeño de LDA regularizado con LDA tradicional en escenarios pequeños de entrenamiento. Para ello utilice los datos 'Datos_Sujeto1_32' y entrene el método cuando sólo 180, 360 ó 720 patrones están disponibles para el entrenamiento. La selección de estos patrones de entrenamiento debe ser tal que las clases se encuentren balanceadas, y deben seleccionarse al azar del conjunto de patrones total. Repita este proceso 10 veces. Para el valor del parámetro de regularización en LDA regularizado utilice el estimador dado por el lemma de Ledoit-Wolf [1].

3.2. Máquinas de soporte vectorial

1. **Utilice las máquinas de soporte vectorial (SVM) para clasificar Imaginería Motora.** Para ello utilice los datos 'DataEEG_MI' (correspondientes a la base de datos "BCI Competition 2008, dataset2a¹"). La misma contiene señales de EEG de dos clases de MI (left and right MI). Los segmentos de EEG fueron extraídos entre 0.5 y 2.5 s luego de aplicada la señal visual. Los datos fueron adquiridos mediante 22 canales de EEG). Filtre cada época de EEG entre 8-30 Hz con un filtro butterworth de orden 5 con corrimiento de fase. Utilice CSP para extraer características discriminativas. Use los primeros 3 pares de filtros espaciales. Transforme las características mediante transformación logarítmica (para que sean linealmente separables). Al finalizar este proceso, debería contar con épocas de EEG de dimensión seis (6). Una vez extraídas las características, utilice SVM lineal para realizar la clasificación. Aplique validación cruzada de 10 particiones. Analice el desempeño del clasificador mediante tasa de acierto, sensibilidad y especificidad.

Recuerde: tanto el proceso de extracción de características como el de aprendizaje deben realizarse con los datos de entrenamiento.

¹<http://www.bbci.de/competition/iv/#dataset2a>

2. Evalúe ahora el método SVM mediante el truco del kernel. Para ello, realice los mismos pasos indicados en el ítem anterior, pero utilizando un kernel gaussiano y uno polinomial. ¿Qué kernel resulta mejor?

3.3. Datos desbalanceados

1. El problema de detección de ERP (P300), tal como se mencionó más arriba, es un problema de clasificación binario pero desbalanceado (1:5). Si los datos presentan más observaciones de una clase, el aprendizaje posiblemente se encuentre *sesgado* hacia la clase mayoritaria, perjudicando a la clase minoritaria. Si bien es posible balancear los datos mediante submuestreo o remuestreo, esto conlleva a eliminar datos útiles o incorporar datos artificiales. Por lo tanto, el método de clasificación debe ser capaz de lidiar con dicho desbalance.

Analice el desempeño de LDA regularizado y SVM lineal para la detección de ERP con desbalance de clase. Utilice los datos 'Datos_Sujeto1_32'. Realice validación cruzada de 10 particiones. Analice el desempeño de los clasificadores utilizando AUC-ROC, AUC-PR y F1. En base a estos resultados, responda: ¿qué algoritmo elegiría para este problema de clasificación?

Referencias

- [1] Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.