A review of Active Learning methods for classification problems

Jorge Bruned Alamán

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

1 Introducción

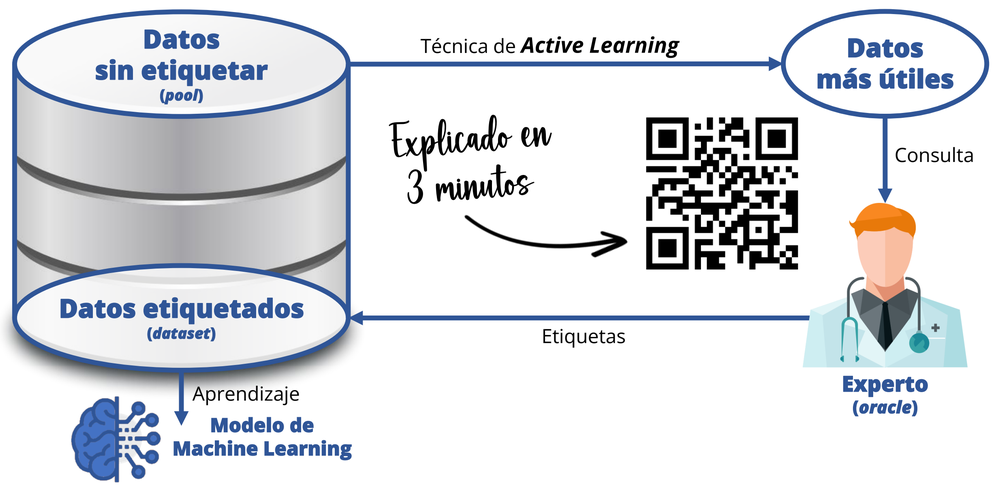
¿Qué es *Active Learning*?

Técnicas de **Machine Learning** utilizadas cuando **no tenemos suficientes etiquetas** para entrenar un modelo.

Permite elegir los **ejemplos más útiles**, que serán **etiquetados** por un **experto** (*oracle*); así **evitamos desperdiciar recursos** al etiquetar ejemplos similares o que aportan poca información.

> Escenarios

- ✓ ***Pool-Based Sampling***: elige los mejores ejemplos para ser etiquetados (nuestro estudio se centra en estas técnicas).



- ✓ ***Membership Query Synthesis***: genera ejemplos artificiales óptimos para ser etiquetados [1].
- ✓ ***Stream-Based Selective Sampling***: decide si etiquetar o no cada ejemplo de forma individual [1].

2 Taxonomía

Técnicas de *Active Learning*

Tras un estudio del **estado del arte**, se han implementado las siguientes técnicas en nuestra propia **librería** de *Active Learning*:

- ✓ ***Random***: baseline.
- ✓ ***Uncertainty-based***: ejemplos para los que el modelo tiene menor confianza.
  - › Least Confidence
  - › Margin Sampling
  - › Confidence Ratio
  - › Entropy Sampling
- ✓ ***Disagreement-based***: ejemplos para los que diferentes modelos se ponen menos de acuerdo.
  - › Query by Committee [2]
  - › Query by Bagging [3]
- ✓ ***Retrain-based***: ejemplos para los que se espera que el modelo mejore más.
  - › Expected Error Reduction
  - › Expected Variance Reduction
  - › Expected Model Change
- ✓ ***Density-based***:
  - › Maximizar distancia a ejemplos etiquetados
  - › Minimizar distancia a ejemplos no etiquetados
- ✓ ***Discriminative AL*** [4]: un modelo intenta distinguir ejemplos etiquetados y no etiquetados para intentar conseguir una muestra representativa.
- ✓ ***Otras técnicas***:
  - › Self-Paced Active Learning (*SPAL*) [5]
  - › Query Informative and Representative Examples (*QUIRE*) [6]
  - › Ensembles: combinación de otras estrategias de esta lista

> Exploración vs. explotación (propuesta)

Nosotros proponemos tratar *Active Learning* como lo que en el fondo es: un **problema de búsqueda**, usando los conceptos de explotación y exploración.

- ✓ ***∈-greedy***: escoger un ejemplo aleatorio con probabilidad ∈ { Valor fijo / Decay rate
- ✓ ***Weighted Sampling***: la probabilidad de escoger cada ejemplo es proporcional a su valor de "utilidad".

3 Metodología

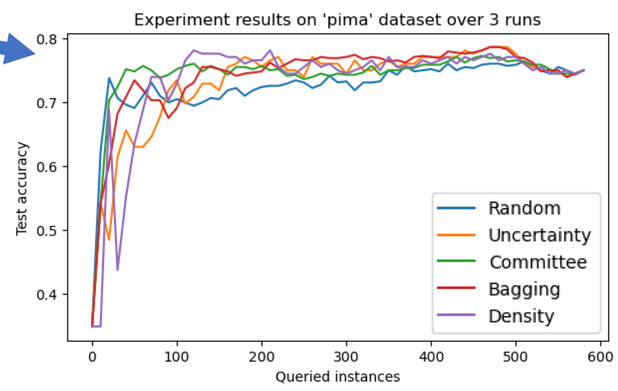
Marco experimental

Utilizamos 56 datasets totalmente etiquetados, pero aplicando una máscara a las etiquetas. De esta forma, el ***oracle*** **conoce las etiquetas** reales *(ground-truth)*.

Métricas: área bajo la curva rendimiento del modelo frente a número de etiquetas

Experimentos

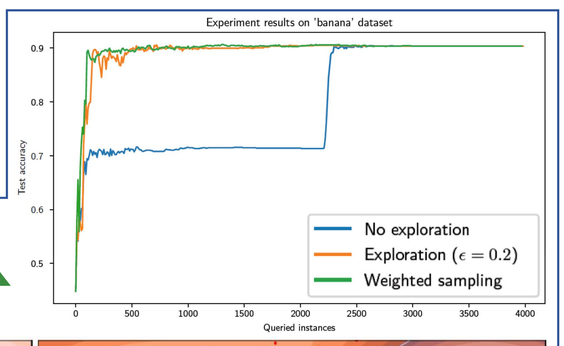
- A Optimizar métodos individualmente
- B Analizar impacto de la exploración
- C Comparar todas las técnicas consideradas



4 Resultados

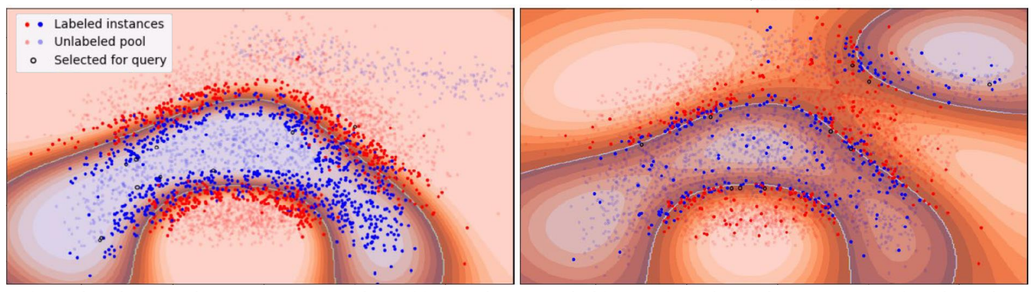
Estudio experimental

A Selección de métricas
- › Confidence ratio (uncertainty)
- › Average *Kullback-Leibler* divergence (disagreement)



B La **exploración** en general es **beneficiosa**

($\epsilon = .2, decay = .05$) **+1.75%** area under the *AL* curve



Ejemplo (dataset *banana*): con exploración se detecta el clúster situado en la zona superior derecha con tan solo ~10% de etiquetas (derecha) mientras que sin ella sigue sin ser detectada con ~40% de etiquetas (izquierda), tal y como se observa en la gráfica performance vs. nº de etiquetas (arriba)

C Comparación de todos los métodos
- › El **muestreo aleatorio** funciona sorprendentemente bien para su sencillez
- › ***Query by Committee*** es el que mejores resultados da en el caso general
- › Para datasets desbalanceados, ***Uncertainty Sampling*** suele ser el ganador

Dataset	Random	Uncertainty	Committee	Bagging	EER	EMC	Density	QUIRE
⋮				⋮				
Average	0.7215	0.7146	**0.7254**	0.7214	0.7234	0.7113	0.6772	0.6872
Median	0.7370	0.7203	**0.7567**	0.7531	0.7556	0.7186	0.7121	0.7122
Avg. rank	3.6364	4.1364	**3.2273**	4.0909	3.2727	4.5909	6.5000	5.8636
Wins	**6**	2	5	3	**6**	2	1	1

5 Conclusiones

Consideramos que hay un **gran potencial** que todavía no se ha explotado al completo. De hecho, **ningún método** de los estudiados **supera consistentemente** al muestreo aleatorio.

Nuestra apuesta son los **métodos combinados** (agregación de múltiples criterios o estrategias), junto a la **exploración**, que ha tenido un **efecto positivo** en la mayoría de los casos.

Además de casos donde partimos de *datasets* con pocas (o sin) etiquetas, una **aplicación** muy interesante sería la **mejora de modelos en producción** gracias a los ejemplos predichos.

> Líneas futuras

- › Considerar nuevas métricas e incorporar **tests estadísticos**.
- › Implementar un mayor número de estrategias de AL.
- › Aplicar en problemas de *deep learning* (ej: visión artificial).
- › **Multi-Instance** Active Learning [7]: mide la "utilidad" de todo el conjunto de ejemplos en lugar de cada uno individualmente.

Referencias

[1] Settles, B. (2009). Active Learning Literature Survey.
[2] Seung, H.S., Opper, M., & Sompolinsky, H. (1992). Query by committee. Annual Conference Computational Learning Theory.
[3] Abe, N., & Mamitsuka, H. (1998). Query Learning Strategies Using Boosting and Bagging. International Conference on Machine Learning.
[4] Guo, Y., & Schuurmans, D. (2007). Discriminative Batch Mode Active Learning. NIPS.
[5] Tang, Y., & Huang, S. (2019). Self-Paced Active Learning: Query the Right Thing at the Right Time. AAAI Conference on Artificial Intelligence.
[6] Huang, S., Jin, R., & Zhou, Z. (2010). Active Learning by Querying Informative and Representative Examples. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 1936-1949.
[7] Settles, B., Craven, M.W., & Ray, S. (2007). Multiple-Instance Active Learning. NIPS.

Jorge Bruned Alamán
contacto@jorgebruned.com

Universidad Pública de Navarra
Grado en Ingeniería Informática | Trabajo de Fin de Grado

Agradecimientos
Mikel Galar Idoate & Neuraptic AI