Práctica: Minería de Flujos de Datos 2024-2025

Master Ciencia de Datos

Universidad de Granada

Práctica: Minería de Flujos de Datos

Miguel García López

Índice

Índice de figuras

Índice de cuadros

1. Cuestiones

1.1. Explica en qué consisten los diferentes modos de evaluación/validación para clasificación en flujos de datos

En clasificación, concretamente en entornos de flujo de datos, los métodos de evaluación difieren de los enfoques estáticos tradicionales debido a la naturaleza dinámica, infinita y potencialmente no estacionaria de los flujos.

Holdout Se toman instantáneas en diferentes momentos durante el entrenamiento del modelo para ver cómo varía la métrica de calidad. Sólo es válido si el conjunto de *test* es similar a los datos actuales (sin *concept drift*) [1].

Test-Then-Train Este enfoque procesa cada nuevo dato en dos fases secuenciales: primero evalúa el modelo (*test*) y luego lo actualiza (*train*). Simula entornos reales de flujos continuos y proporciona métricas en tiempo real, como precisión acumulada.

Prequencial Variante de *Test-Then-Train* que calcula métricas en ventanas deslizantes o bloques. Utiliza dos enfoques: ventanas fijas (evalúa últimos nn datos, como 1000 ejemplos) o ventanas adaptativas (ajusta el tamaño según detección de *drift*, como el algoritmo *ADWIN* [2]). Su ventaja principal es reducir el sesgo hacia datos antiguos. En [3], se aplicó en flujos financieros para medir la adaptabilidad de modelos ante cambios de mercado.

Interleaved Validation Adaptación de la validación cruzada tradicional: divide el flujo en bloques temporales y los rota para entrenamiento y prueba. Este método es útil para evaluar robustez frente a *drift*. En [4], se empleó para comparar algoritmos como *VFDT* y *Hoeffding Adaptive Tree* en presencia de cambios sintéticos en la distribución.

Ventanas Deslizantes (Sliding Windows) Evalúa el modelo solo en datos recientes. Las ventanas pueden ser fijas (mantienen tamaño constante, como los últimos 10000 ejemplos) o adaptativas (ajustan dinámicamente el tamaño usando umbrales de error [5]). Un ejemplo clásico es *VFDT* [6], que usa ventanas para limitar el uso de memoria en flujos infinitos, descartando datos obsoletos.

1.2. Describe tres algoritmos de clasificación en flujos de datos y compara ventajas/desventajas

El primer algoritmo y uno de los más usados es el **VFDT** (*Very Fast Decision Tree*) o Árbol de *Hoeffding*. El **VFDT**, propuesto por *Domingos* y *Hulten* (2000), es un algoritmo incremental que construye árboles de decisión utilizando el *Hoeffding bound* (HB), un límite estadístico que garantiza con alta probabilidad que la mejor división en un nodo, basada en una muestra de datos, será la misma que si se usara el flujo completo. Opera en

tiempo constante por muestra y memoria limitada. La cota *Hoeffding* se describe como:

$$HB = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$$

Donde R es el rango de clases (diferencia máxima posible en las métricas de división, como ganancia de información), n el número de muestras en el nodo, y δ la probabilidad de error [6].

Usa la cota de *Hoeffding* para garantizar que, con alta probabilidad, la mejor elección de división con una muestra será la misma que si se usara todo el flujo de datos. Esto permite hacer divisiones rápidamente sin necesidad de almacenar todos los datos históricos.

2. Bibliografía

[1] J. Casillas, *Minería en Flujo de Datos*, Máster en Ciencias de Datos e Ingeniería de Computadores, Universidad de Granada, 2025. URL: `http://decsai.ugr.es/~casillas`.

[2] A. Bifet y R. Gavaldà, "Learning from Time-Changing Data with Adaptive Windowing," en *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007, págs. 443-448.

[3] J. Gama, *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC, 2010, ISBN: 978-1439826119.

[4] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy y A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys*, vol. 46, n.º 4, págs. 1-37, 2014.

[5] A. Bifet, G. Holmes, B. Pfahringer y R. Gavaldà, "Adaptive Hoeffding Trees for Learning from Data Streams," en *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)*, 2009, págs. 139-147.

[6] P. Domingos y G. Hulten, "Mining High-Speed Data Streams," en *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, págs. 71-80.