DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

SERIES TEMPORALES Y MINERÍA DE FLUJOS DE DATOS

E.T.S. de Ingenierías Informática y de Telecomunicación

Trabajo Autónomo II

Minería de Flujos de Datos

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL
UNIVERSIDAD DE GRANADA

1. Introducción

El objetivo de este guion es que el alumno conozca el software de análisis de flujos de datos MOA y realice una discusión acerca de los resultados de aplicar diferentes técnicas para resolver problemas de clasificación. Para ello, se propondrán ejercicios que el alumno deberá resolver y cuyos resultados deberán ser analizados y discutidos.

2. Descripción del trabajo a realizar

Se proponen diversas tareas para que se resuelvan por parte del alumno, a fin de evaluar modelos estáticos y dinámicos, en minería de flujos de datos sin y con desvío de concepto:

2.1. Entrenamiento offline (estacionario) y evaluación posterior.

1. Entrenar un clasificador HoeffdingTree offline (estacionario, aprender modelo únicamente), sobre un total de 1.000.000 de instancias procedentes de un flujo obtenido por el generador WaveFormGenerator con semilla aleatoria igual a 2. Evaluar posteriormente (sólo evaluación) con 1.000.000 de instancias generadas por el mismo tipo de generador, con semilla aleatoria igual a 4. Repita el proceso varias veces con la misma semilla en evaluación y diferentes semillas en entrenamiento, para crear una población de resultados. Anotar como resultados los valores de porcentajes de aciertos en la clasificación y estadístico Kappa.
2. Repetir el paso anterior, sustituyendo el clasificador por HoeffdingTree adaptativo.
3. Responda a la pregunta: ¿Cree que algún clasificador es significativamente mejor que el otro en este tipo de problemas? Razone su respuesta.

2.2. Entrenamiento online.

1. Entrenar un clasificador HoeffdingTree online, mediante el método Interleaved Test-Then-Train, sobre un total de 1.000.000 de instancias procedentes de un flujo obtenido por el generador WaveFormGenerator con semilla aleatoria igual a 2, con una frecuencia de muestreo igual a 10.000. Pruebe con otras semillas aleatorias para crear una población de resultados. Anotar los valores de porcentajes de aciertos en la clasificación y estadístico Kappa.
2. Repetir el paso anterior, sustituyendo el clasificador por HoeffdingTree adaptativo.
3. Responda a la pregunta: ¿Cree que algún clasificador es mejor que el otro en este tipo de problemas? Razone su respuesta.

2.3. Entrenamiento online en datos con concept drift.

1. Entrenar un clasificador HoeffdingTree online, mediante el método Interleaved Test-Then-Train, sobre un total de 2.000.000 de instancias muestreadas con una frecuencia de 100.000, sobre datos procedentes de un generador de flujos RandomRBFGeneratorDrift, con semilla aleatorio igual a 1 para generación de modelos y de instancias, generando 2 clases, 7 atributos, 3 centroides en el modelo, drift en todos los centroides y velocidad de cambio igual a 0.001. Pruebe con otras semillas aleatorias. Anotar los valores de porcentajes de aciertos en la clasificación y estadístico Kappa. Compruebe la evolución de la curva de aciertos en la GUI de MOA.
2. Repetir el paso anterior, sustituyendo el clasificador por HoeffdingTree adaptativo.
3. Responda a la pregunta: ¿Cree que algún clasificador es mejor que el otro en este tipo de problemas? Razone su respuesta.

2.4. Entrenamiento online en datos con concept drift, incluyendo mecanismos para olvidar instancias pasadas.

1. Repita la experimentación del apartado anterior, cambiando el método de evaluación “Interleaved Test-Then-Train” por el método de evaluación “Prequential”, con una ventana deslizante de tamaño 1.000.
2. ¿Qué efecto se nota en ambos clasificadores? ¿A qué es debido? Justifique los cambios relevantes en los resultados de los clasificadores.

2.5. Entrenamiento online en datos con concept drift, incluyendo mecanismos para reinicializar modelos tras la detección de cambios de concepto.

1. Repita la experimentación del apartado 2.3, cambiando el modelo (learner) a un clasificador simple basado en reemplazar el clasificador actual cuando se detecta un cambio de concepto (SingleClassifierDrift). Como detector de cambio de concepto, usar el método DDM con sus parámetros por defecto. Como modelo a aprender, usar un clasificador HoeffdingTree.
2. Repita el paso anterior cambiando el clasificador HoeffdingTree por un clasificador HoeffdingTree adaptativo.
3. Responda a la siguiente pregunta: ¿Qué diferencias se producen entre los métodos de los apartados 2.3, 2.4 y 2.5? Explique similitudes y diferencias entre las diferentes metodologías, y discuta los resultados obtenidos por cada una de ellas en el flujo de datos propuesto.

3. Condiciones de entrega

La entrega se realizará mediante la presentación en formato electrónico de una memoria de teoría y prácticas, junto con ficheros separados que resuelvan cada problema. Todos los ficheros se adjuntarán en un .ZIP que se entregará mediante la plataforma docente de la asignatura. La práctica contribuirá a la calificación final de la asignatura en **4 puntos**, divididos entre **2 puntos** para la parte de teoría y **2 puntos** para la parte práctica.

La memoria tendrá **dos partes claramente diferenciadas**:

- Parte teórica (2 puntos), donde se resolverán cuestiones básicas de minería de flujo de datos y se describirán los conceptos teóricos que se han utilizado en las prácticas:
  - (1 punto) Responda a las 12 preguntas tipo test del anexo. Cada respuesta incorrecta restará un tercio (si hay cuatro opciones) o una unidad (si tiene dos opciones) del valor de cada pregunta (un doceavo).
  - (0.5 puntos) Explique el problema de clasificación, los clasificadores utilizados en los experimentos de la sección 2, y en qué consisten los diferentes modos de evaluación/validación en flujos de datos. Desarrolle con suficiente detalle este apartado.
  - (0.5 puntos) Explique en qué consiste el problema de *concept drift* y qué técnicas conoce para resolverlo en clasificación. Desarrolle con suficiente detalle este apartado.
- Parte práctica (2 puntos), donde se describirán los comandos utilizados para generar las soluciones de cada apartado, indicando porqué si han utilizado los parámetros usados y sus valores. Los resultados de la precisión de la clasificación (% aciertos y estadístico Kappa) se presentarán en una tabla. El alumno aportará una discusión justificada, basándose en estos datos, sobre el comportamiento de cada método en los apartados de los problemas planteados en la sección 2.

La memoria también deberá contener una portada con la siguiente información:

- Nombre del alumno
- E-mail del alumno
- Nombre de la asignatura
- Nombre del Máster
- Texto “Trabajo autónomo II: Minería de Flujo de Datos”

Anexo: Teoría – Minería de Flujos de Datos. Curso 2018-2019.

Apellidos y nombre: _____

1. El aprendizaje incremental es útil cuando...
   - ☐ se trabaja con datos no estacionarios
   - ☐ aumenta el tiempo de respuesta
   - ☐ se aprende sobre otro algoritmo
   - ☐ se quiere ganar eficiencia
2. La minería de flujo de datos se considera cuando...
   - ☐ el problema genera datos continuamente
   - ☐ los datos son estacionarios
   - ☐ se quiere mejorar principalmente la eficiencia
   - ☐ se quiere mejorar un modelo previamente aprendido
3. La cota de Hoeffding sirve para saber...
   - ☐ si la información recibida es fiable
   - ☐ cuándo hay suficientes datos para una estimación fiable
   - ☐ qué fiabilidad tienen los datos
   - ☐ qué precisión se puede alcanzar
4. ¿Que características de clusters mantiene el algoritmo BIRCH?
   - ☐ Suma lineal y número de objetos
   - ☐ Tiempo, suma lineal y tamaño
   - ☐ Tiempo, suma lineal y suma cuadrática
   - ☐ Suma lineal, suma cuadrática y número de objetos
5. ¿El algoritmo Stream maneja concept drift?
   - ☐ Sí
   - ☐ No
6. ¿Qué es concept drift?
   - ☐ Cambios en el número de clases
   - ☐ Variaciones en los parámetros del algoritmo
   - ☐ Variaciones en los tipos de variables
   - ☐ Cambios en la dinámica del problema
7. ¿Cómo gestiona CVFDT el concept drift?
   - ☐ Sustituye el atributo del nodo por el segundo mejor
   - ☐ Crea un cluster alternativo
   - ☐ Detecta el cambio de concepto
   - ☐ Mantiene árboles alternativos
8. ¿Por qué es útil el ensemble learning en concept drift?
   - ☐ Porque es robusto frente a cambios de párametros
   - ☐ Porque aprovecha la diversidad que se genera en los cambios
   - ☐ Porque matienen muchos modelos simultáneamente
   - ☐ Porque hace un muestreo selectivo de los datos
9. ¿Cuál es más eficiente entre DDM y ADWIN?
   - ☐ DDM es más eficiente
   - ☐ Los dos son muy ineficientes
   - ☐ ADWIN es más eficiente
   - ☐ Los dos son similares
10. ¿Por qué es controvertida la clasificación en flujo de datos?
    - ☐ Porque se requiere al oráculo por siempre
    - ☐ Porque el etiquetado es costoso
    - ☐ Porque la clase puede cambiar
    - ☐ Porque el oráculo puede cambiar su valoración
11. ¿Cómo gestiona ClueStream el concept drift?
    - ☐ Expande los microclusters cuando llega un cambio de concepto
    - ☐ Reduce las fronteras inter-microclusters si es necesario
    - ☐ Mantiene información sobre el tiempo
    - ☐ Reduce el tamaño de los microclusters cada cierto tiempo
12. ¿Por qué es complejo generar reglas de asociación en flujo de datos?
    - ☐ Porque los frequent itemsets no pueden ser dinámicos
    - ☐ Porque las reglas se evalúan sobre el histórico de datos
    - ☐ Porque para calcular la confianza se requieren muchos datos
    - ☐ Porque no maneja bien variables continuas