TRABAJO FIN DE MÁSTER

Máster DATCOM: Ciencia de Datos

Árboles de clasificación monotónica sobre flujos de datos.

Autor
Carlos Manuel Sequí Sánchez(alumno)

Directores
Salvador García López(tutor)

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, septiembre de 2019

Árboles de clasificación monotónica sobre flujos de datos.

Autor
Carlos Manuel Sequí Sánchez

Directores
Salvador García López

Departamento de Ciencias de la Computación e I.A.

Árboles de clasificación monotónica sobre flujos de datos.

Carlos Manuel Sequí Sánchez(alumno)

Palabras clave: Monotonic, data streams, decision tree, Hoeffding tree, classification, MOA

Resumen

En la presente documentación se tratarán diversas temáticas relativas a la ciencia de datos tales como son, los árboles de decisión, los flujos de datos y la clasificación monotónica.

En la primera parte del documento se pondrá al lector en el contexto del problema mediante la explicación de manera detallada de estas técnicas, así como ejemplos de uso, ventajas e inconvenientes, etc.

Posteriormente se procederá a la exposición de la propuesta con el fin de hacer entender al lector que consiste en una técnica novedosa en la materia. Esta consiste en una adaptación a un algoritmo existente de árboles de decisión para flujos de datos que además, en este caso, posean restricciones monotónicas para hacer que los modelos aprendidos a partir de los datos sean más fieles a la realidad.

Tras la definición de la propuesta se describirán los detalles de los experimentos implementados para la realización de la propuesta, es decir, el **marco de trabajo**, el cual incluye una descripción de los conjuntos de datos empleados y medidas para la comparativa de algoritmos, **resultados** de los experimentos y, finalmente, **análisis** de estos.

Finalmente se finalizará el documento con una serie de conclusiones y posibles trabajos futuros.

Monotonic classification trees on data streams.

Carlos Manuel Sequí Sánchez(student)

Keywords: Monotonic, data streams, decision tree, Hoeffding tree, classification, MOA

Abstract

This documentation will cover various topics related to data science such as decision trees, data flows and monotonic classification.

In the first part of the document the reader will be placed in the context of the problem by explaining in detail these techniques, as well as examples of use, advantages and disadvantages, etc.

Subsequently, the proposal will be presented in order to make the reader understand that it consists of a new technique in the field. This consists of an adaptation to an existing decision tree algorithm for data flows that also, in this case, have monotonic restrictions to make the models learned from the data more faithful to reality.

After defining the proposal, the details of the experiments implemented for the realization of the proposal will be described, that is, the **framework**, which includes a description of the data sets used and measures for the comparison of algorithms, **results** of the experiments and, finally, **analysis** of these.

Finally, the document will be finished with a series of conclusions and possible future work.

Yo, **Carlos Manuel Sequí Sánchez**, alumno de la titulación Máster DATCOM: Ciencia de Datosde la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación**, con DNI 20486926K, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Máster en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Carlos Manuel Sequí Sánchez

Granada a 1 de septiembre de 2019.

D. **Salvador García López(tutor1)**, Profesor del Departamento de Ciencias de la Computación e I.A. de la Universidad de Granada.

Informa:

Que el presente trabajo, titulado ***Árboles de clasificación monotónica sobre flujos de datos.***, ha sido realizado bajo su supervisión por **Carlos Manuel Sequí Sánchez**, y autoriza la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expide y firma el presente informe en Granada a 1 de Septiembre de 2019 .

El director:

Salvador García López

Agradecimientos

Llegado a este punto, agradezco la paciencia e interés puesto en mi aprendizaje a todos los profesores que han formado parte, durante todo este año, de poner en mis manos la semilla de conocimiento que me servirá para lanzarme al mundo profesional, así como a mi tutor Salvador García López, quien se ha encargado de ayudarme y supervisar este TFG. Agradezco a toda mi familia y, con mayor énfasis a mis padres y a mi hermano, el interés y el apoyo que me han ofrecido desde el primer momento, aunque no entiendan del todo las "letras raras" en la pantalla de mi ordenador cuando trabajo, o que no nos enseñen a "hackear" cosas. Por último, agradezco el haber prolongado el contacto en el ámbito académico con los amigos que hice durante el grado, con buena compañía todo ha sido más sencillo, ya sabéis.

Índice general

Índice de figuras

Capítulo 1

Introducción y objetivos

1.1. Introducción

¿Qué es lo primero que se nos viene a la cabeza cuando escuchamos la palabra **informática**?¿Un ordenador?¿Datos de infinidad de temas y tipos almacenados?¿Programas para facilitar el uso de esos datos?¿Internet?

Si mezclamos estas ideas que nos surgen de forma repentina al pensar en dicha palabra, podemos extraer una definición apropiada con facilidad. Una de tantas definiciones para ello, es que la informática se trata de la ciencia que se encarga del almacenamiento, procesamiento y transferencia de información en formato digital a través de un sistema físico con capacidades computacionales (hardware). Mediante dicho **hardware**, podemos crear programas (**software**) que faciliten la automatización del proceso de almacenamiento, procesamiento y transferencia de información citados.
¿Quién no ha utilizado un ordenador en alguna ocasión para jugar a su juego favorito, imprimir un documento, buscar información por Internet o para ver un vídeo en Youtube de su cantante favorito?
Todas estas acciones realizadas con un ordenador (por ejemplo) resultan ser obras de personas con conocimientos específicos en informática necesarios para la creación de estos complejos sistemas dedicados al tratamiento de información de una u otra forma distinta.

Si nos adentramos un pequeño paso más hacia aspectos técnicos de la informática, podremos hablar del concepto de **programación** que, a rasgos muy generales, es el proceso mediante el cual una persona (programador) crea una especie de receta con instrucciones (código fuente) que el ordenador deberá seguir paso a paso para la realización de la tarea deseada por el programador y, es de esta manera la forma de la que se crea un programa.

Una vez tenemos la forma de crear procesos automáticos para facilitar a las personas el tratamiento de información, surgen a lo largo de la historia de la informática infinidad de aplicaciones con la que explotar estas capacidades computacionales para sacar provecho económico (empresas privadas), moral (sistemas de vigilancia), académico (clases online) y un largo etcétera de ámbitos de aplicación distintos.
Es aquí donde introduzco el concepto clave para la descripción del tema a tratar en este proyecto: **la Ciencia de Datos**, que en términos muy generales, es un campo interdisciplinar que, a través del uso de ciencias como las matemáticas, estadística e informática, se dedica al análisis y procesamiento de datos para la extracción de información útil de conjuntos de datos(Data Mining), toma de decisiones con respecto a dichos análisis(Estadística inferencial), y automatización de dichos procesos de toma de decisiones(Machine Learning).

Como podemos imaginar, poseer conocimientos sobre esta ciencia, supone grandes beneficios para, por ejemplo, empresas que tienen en sus dominios ingentes cantidades de datos sobre sus productos o clientes que los consumen, ya que puede aportarle información para la mejora de la calidad de los productos, mejora de servicios para satisfacer a los clientes, conocimientos sobre el tipo de cliente al que se orienta el producto para crear estrategias de marketing...en conclusión, infinidad de aplicaciones que permiten analizar datos, extraer conclusiones y crear programas automáticos de toma de decisión que un ser humano sería incapaz de realizar en tan poco tiempo y de una forma tan efectiva. Ejemplos de uso de esta ciencia son los de Uber, con su sistema de optimización de ruta en las ciudades, o Amazon, con su sistema de recomendación basado en la predicción de necesidades de sus clientes en función de sus historiales de compra.

Para la realización de estos procesos de análisis y procesamiento de datos es indispensable la utilización de algoritmos (esas recetas de las que hablábamos antes) y técnicas complejas de estructuración y análisis de los datos que poseemos en bruto recolectados de alguna manera, que faciliten dichas tareas tanto en tiempo de ejecución como en precisión en la toma de decisiones. Una de estas técnicas de aprendizaje a partir de datos son los llamados **árboles de decisión**, llamados así porque su estructura (similar a la de un árbol al revés con raíz, ramas y hojas) permite a una computadora "**aprender**" conocimientos sobre un conjunto de datos organizado y **tomar decisión** en base a ese conocimiento adquirido mediante los ejemplos analizados.

Los datos tratados por estas astutas técnicas de análisis no siempre permanecen almacenados y ordenados previo análisis, si no que pueden provenir de fuentes que generan ingentes cantidades de datos de forma periódica y sin pausa, los cuales han de ser tratados de forma rápida y constante para poder obtener una buena representación de su comportamiento y poder generar una toma de decisión correcta. Estos son los llamados **flujos de datos o data streaming**.

Es bien sabido también que, toda ayuda que el ser humano pueda poner de su parte para la mejora de los algoritmos de análisis y toma de decisiones, será bien recibida por cualquiera de las técnicas utilizadas en la Ciencia de Datos. Es esta la idea de hacer uso de las llamadas **restricciones monotónicas** de las cuales hablaremos durante todo el documento. Consiste básicamente en aportar información subyacente a un tipo problema específico para ayudar al algoritmo a alcanzar mejores resultados.

1.2. Motivación

Tras el estudio a lo largo del máster de todo lo detallado en el apartado anterior y tras ver ejemplos y ejemplos de uso de los citados árboles de decisión, técnicas de data streaming y técnicas de uso de restricciones monotónicas, no he sido capaz de encontrar artículos ni ejemplos donde se haga uso de estos tres elementos de forma conjunta con el fin de obtener mejores resultados sobre flujos de datos con capacidades de obedecer restricciones monotónicas mediante el uso de árboles de decisión.

En particular, con respecto a las técnicas de flujos de datos, se puede observar en la literatura la inexistencia del tratamiento de estos usando árboles de decisión con restricciones de monotonía.

Como bien he dicho anteriormente, toda ayuda que el ser humano pueda aportar a los algoritmos que se encargan de la resolución de cualquier problema, será bien recibida, por tanto, pienso que incluir conocimiento mediante restricciones de monotonía sobre los datos a un algoritmo que utiliza árboles de decisión para resolver problemas de data streaming hará que se obtengan mejores resultados.

De esta manera, la motivación principal es aportar a la comunidad científica un novedoso tratamiento de algoritmos ya existentes para mejorar la calidad de estos en problemas específicos relativos a la Ciencia de Datos.

1.3. Objetivos y estructuración de la memoria

El objetivo inicial de este documento es dotar al lector de los conocimientos y el contexto necesarios sobre las distintas técnicas a utilizar para que logre entender la finalidad de la propuesta (Capítulos 2 y 3). En esta primera parte del documento se explicarán en profundidad los elementos ya mencionados a formar parte de objeto de estudio:

- Clasificación sobre flujos de datos.
- Clasificación con restricciones monotónicas.
- Árboles de decisión y su uso con data streams y restricciones monotónicas (por separado, evidentemente).

Una vez hayamos dotado al lector de la contextualización del problema, el siguiente objetivo (Capítulo 4) es la descripción detallada de la propuesta con el fin de hacer entender las ideas necesarias para la creación del nuevo algoritmo, los pasos a seguir para lograrlo y los resultados esperados con ello, seguido de la explicación del software utilizado para tal propósito (Capítulo 5), así como su uso.

Finalmente, aparecerá en este documento una exposición de los experimentos realizados con la propuesta (Capítulo 6), así como la descripción del marco de trabajo en el que lo situaremos, además de comparaciones de sus resultados con los de los algoritmos que citaremos más adelante, con el propósito de observar si la propuesta cumple el cometido de resultar ser mejor en los aspectos que deseemos.

Acompañado de estas comparaciones y para finalizar el documento, presentaremos también una serie de conclusiones y trabajos futuros a desarrollar para continuar con esta línea de trabajo (Capítulo 7).

Capítulo 2

Contexto en problemas

2.1. Clasificación sobre flujos de datos

Los sistemas tradicionales basados en el uso de memoria, entrenados de una forma fija mediante conjuntos de entrenamiento y los cuales generan modelos estáticos, no están preparados para procesar los datos altamente detallados disponibles en procesos como, por ejemplo, el continuo análisis de datos generados por los sensores de una máquina que trabaja sin descanso, lo cual crea una gran cantidad de datos que ha de ser procesada de forma rápida con el fin de generar modelos predictivos consistentes que se adapten a situaciones cambiantes y puedan reaccionar de forma rápida y eficaz a dichos cambios.

El Machine Learning extrae conocimiento en forma de modelos y patrones de unos datos de naturaleza cambiante. Hoy en día la generación de datos, gracias a las capacidades tecnológicas de las que disfrutamos, se produce a altas velocidades, tanto es así, que se pone, en cuanto a velocidad, por delante del procesamiento de dichos datos, lo cual quiere decir que generamos datos a mayor velocidad de lo que las capacidades computacionales que tenemos ahora mismo nos permiten procesarlos.
Desde este punto de vista, en estos casos conviene modelar los datos como flujos de datos transitorios en lugar de como tablas de datos persistentes.

	Databases	Data streams
Data access	Random	Sequential
Number of passes	Multiple	Single
Processing time	Unlimited	Restricted
Available memory	Unlimited	Fixed
Result	Accurate	Approximate
Distributed	No	Yes

Figura 2.1: Resumen entre las diferencias principales entre el procesamiento estándar de una base de datos y el procesamiento de flujo de datos. [10]

2.1.1. Tipos de algoritmos para flujos de datos

Existen dos tipos distintos de algoritmos que trabajan sobre flujos de datos:

- **Insert-only model**: donde los datos entran al sistema de forma secuencial.
- **Insert-delete model**: donde los elementos que entran pueden ser eliminados o actualizados.

Desde el punto de vista de los sistemas de control de flujo de datos(DSMS), existen varios problemas que requieren técnicas de procesamiento no exactas para evaluar el flujo continuo de datos que requieren una cantidad ilimitada de memoria.

Estos algoritmos de procesamiento flujos de datos producen soluciones aproximadas dentro de un rango de error admisible para ciertas aplicaciones, con una alta probabilidad, relajando así las restricciones a la hora de obtener una solución exacta.

Los sistemas de control de flujos de datos han desarrollado un conjunto de técnicas que almacenan resúmenes de datos compactos suficientes para resolver consultas. Estas aproximaciones requieren un equilibrio entre el accuracy y la cantidad de memoria usada para almacenar los resúmenes, con una restricción adicional de tiempo de procesado de los datos.

2.1.2. Aproximación y aleatorización

Dentro del marco del data streaming, como ya hemos dicho, está permitido ofrecer respuestas aproximadas dentro de un pequeño rango de error

(ϵ), con una pequeña probabilidad de fallo (δ) para obtener respuestas con una probabilidad de que 1-δ se encuentre en el intervalo de radio ϵ.

Los algoritmos que usan estas aproximacióon y aleatorización son referidos por dichos (ϵ,δ).
la idea consiste básicamente en mapear cada espacio grande de entrada en una sinopsis pequeña.
La aproximación y randomización han sido usadas en solventar problemas como minería de reglas de asociación, items frecuentes, k-means...

2.1.3. Ventanas de tiempo

Para la realización del cómputo estadístico referente al modelo de flujos, no nos interesa el total de los datos existentes, si no los más recientes, entendiendo que son los que mejor explican la situación a la que nos enfrentamos y pudiendo, de esta forma, deshacernos de grandes cantidades de datos que no nos son útiles.

Las técnicas más simples para este tipo de tratamiento de datos, utilizan una ventana deslizante de tamaño fijo, con un funcionmiento FIFO (first in fist out).

Definimos dos tipos de ventana deslizante:

- **Basada en secuencia**: donde el tamaño de ventana queda definido por el número de observaciones del data set (tamaño fijo o variable en el tiempo).
- **Basada en marca de tiempo**: donde el tamaño de ventana está definido en términos de duración. Una ventana de este tipo de tamaño t consiste en todos los elementos cuya marca de tiempo se situa dentro del intervalo de tiempo t del actual periodo de tiempo.

El hecho de monitorizar, analizar y extraer conocimiento de flujos de datos de alta velocidad, puede hacer que existan diversos niveles de **granularidad** a la hora de almacenar los datos. Conforme más antiguos son los datos que disponemos, mayor granularidad requeriremos en la información (es decir, menor precisión). Cuanto más reciente sean los datos, el grano ha de ser más fino, ya que requerimos más precisión al tratarlos debido a que son más importantes (este es llamado el **modelo de ventana de tiempo inclinado**).

Ejemplo de algoritmo de ventana de tiempo

AdWin-ADaptive sliding WINdow: mantiene una ventana variable con respecto a los items recientemente vistos con la propiedad de que la ven-

tana tiene un tamaño maximal estadísticamente consistente con la hipótesis de que no haya habido un cambio en la media del valor dentro de la ventana. Un fragmento viejo de la ventana se desecha si hay alguna evidencia de que tiene un valor distinto al del resto de la ventana.

2.1.4. Muestreo

El sampling (o muestreo) consiste en la selección del subconjunto de datos a analizar en intervalos periódicos, utilizado para calcular estadísticas del flujo (valores esperados).
Este tipo de técnicas reduce la cantidad de datos a procesar, por tanto, el coste computacional.
Como contra a su uso, podemos decir que pueden ser una fuente de errores, por ejemplo, en aplicaciones dedicadas a la detección de valores extremos o anomalías, ya que, a la hora de realizar el sampling podemos estar eliminando dichas instancias. El problema principal es obtener una muestra representativa.

Técnicas de muestreo:

- **Random sampling**: muestreo aleatorio de los datos (todas las instancias con la misma probabilidad de ser escogidas).
- **Reservoir sampling**: consiste en mantener una muestra de tamaño K de reserva. A medida que fluyen los datos, cada nuevo elemento tiene una probabilidad k/n (donde n son los datos visualizados hasta el momento) de reemplazar un antiguo dato.
- **Load shedding**: elimina secuencias del flujo de datos cuando se producen cuellos de botellas en las capacidades de procesamiento.

2.1.5. Sinopsis, bocetos y resúmenes

A continuación describimos tres métodos de compactación de información para la generación de modelos sobre los ya comentados conjuntos de datos reducidos para data streaming:

- **Sinopsis**: estructuras de datos compactas que resumen datos para su posterior consulta.
- **Data sketching**: herramienta de reducción de dimensionalidad. Usa proyecciones aleatorias de datos con cierta dimensión d a un espacio de cierto conjunto de dimensiones.

- **Data stream summary (by Cirnide and Muthukrishnan)**: usado para aproximaciones (ϵ,δ) para resolver consultas de rango, consultas puntuales y consultas innerproduct.

2.1.6. Problemas en el aprendizaje sobre data streams

El objetivo de la minería de datos es la habilidad de mantener de forma permanente un modelo de decisión preciso. Este problema requiere algoritmos que se adapten a los datos conforme esten disponibles para poder aprender de ellos. Además, los datos desactualizados han de ser olvidados para dejar de tenerlos en cuenta a la hora de crear el modelo, cosa que ha de ocurir en la presencia de información con una distribución no estacionaria presente. El aprendizaje en flujos de datos requiere por tanto algoritmos incrementales de aprendizaje que tengan en cuenta el llamado çoncept drift".

La solución a estos problemas requieren nuevas técnicas de muestreo y randomización, y nuevos algoritmos aproximados, incrementales y decrementales. Algunas propiedades deseables para algoritmos de flujos de datos:

- **Incrementalidad**
- **Aprendizaje online**
- **Tiempo constante de procesado de cada ejemplo**
- **Un solo escaneo sobre el conjunto de datos de training**
- **Tener en cuenta el concept drift**

Los algoritmos de aprendizaje incrementales y decrementales requieren una permanente actualización del modelo de decisión conforme llegan datos nuevos. Esta habilidad de actualizar el modelo mediante las propiedades de los nuevos datos es importante, pero no suficiente, ya que también es necesaria la habilidad de olvidar información anticuada para dar un giro en el aprendizaje realizado, dejando de tener en cuenta los items antiguos: decremental learning.

Evidentemente existe una balanza entre la ganancia en rendimiento ofrecido por el algoritmo y la manutención de la característica de actualización de este, lo que hace que el cómputo realizado por el algoritmo sea más complejo. Ante esta balanza, con el fin de no acrecentar el cómputo, haciendo que el algoritmo decida de forma dinámica qué información borrar y cuál no, surge la ya comentada técnica de ventana deslizante.

De forma general, es complicado asumir que, en el manejo de flujos de datos durante un largo tiempo, estemos tratando con datos acordes a una

distribución de probabilidad estacionaria. En sistemas complejos y en largos periodos de tiempo, debemos esperar cambios en la distribución de los items.

Una aproximación natural para estas tareas incrementales son los algoritmos de aprendizaje adaptativo, algoritmos incrementales que tienen en cuenta el concept drift. El concept drift en sí, se refiere al cambio de concepto que sufren los datos a la largo del tiempo, cada vez con cierta permanencia mínima. Hay algoritmos que implementan el olvido de información antigua teniendo en cuenta este cambio de concepto, lo que los hace mucho más precisos que los propios algoritmos que realizan la eliminación de información en forma de tamaño de ventana prefijado.
Con el uso de los algoritmos de detección de concept drift podemos averiguar cuando y por qué ha cambiado el comportamiento del flujo de datos.
Estos algoritmos no poseen la información de mundo cerrado de la que disponen los algoritmos convencionales para el tratamiento de datos estáticos, si no que han de ser capaces de adaptarse a un mundo abierto cambiante de datos para diferenciar entre cambio de concepto y ruido en los datos.

A la hora de evaluar los resultados en el contexto de los flujos de datos, es interesante tener en cuenta la evolución del acierto de nuestro algoritmo a lo largo del tiempo con los cambios de concepto acaecidos.

2.1.7. Requisitos y funcionamiento de flujos de datos

A modo de resumen, planteamos los requisitos primordiales para la clasificación de data streams de la siguiente forma:

1. Procesamiento de un ejemplo en cada instante de tiempo e inspección de este tan solo una vez.
2. Limitado uso de memoria.
3. Trabajo en un tiempo limitado.
4. Estar listo para predecir en cualquier momento.

De la misma forma, describimos el ciclo de clasificación para flujos de datos

1. El algoritmo toma el siguiente ejemplo del flujo
2. El algoritmo procesa el ejemplo actualizando sus estructuras de datos. En este punto no se ha de exceder los límites de memoria y ha de ser lo más rápido posible.
3. El algoritmo está listo para aceptar el siguiente ejemplo.

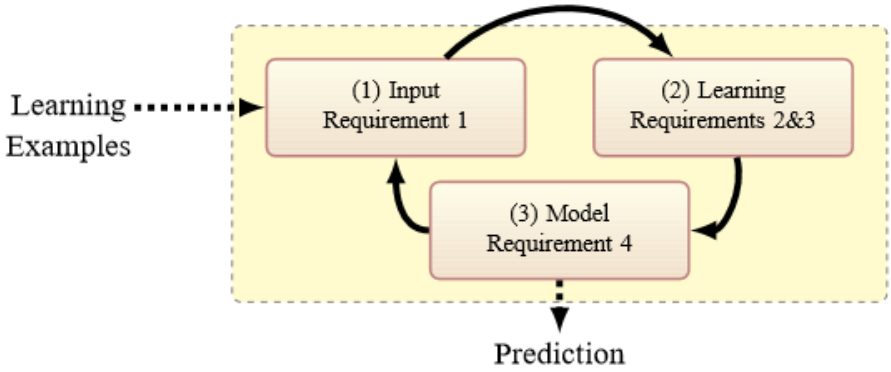

Figura 2.2: Ciclo de clasificación para flujos de datos junto con los requisitos utilizados en cada paso de los descritos anteriormente.

Para el procedimiento de evaluación de los algoritmos de aprendizaje, mientras que los modelos batch tradicionales utilizan un conjunto de datos de train y otro de test de reserva (holdout) para realizar dicha evaluación, en los algoritmos dedicados al data streaming se utiliza el método Interleaved Test-Then-Train o Prequential, mediante el cual cada una de las instancias que se reciben se utilizan como instancia de test para, posteriormente, usarla como nuevo dato de aprendizaje (train). De esta forma no es necesario mantener un conjunto de datos de reserva exclusivo para validar, haciendo que el uso de los datos disponibles sea máximo, además de que ayuda a crear una representación más visual de la evolución de la precisión del algoritmo a lo largo del tiempo.

2.2. Clasificación ordinal y monotónica

Comenzamos este apartado definiendo primeramente un par de conceptos básicos para adentrarnos de forma correcta en el significado de la clasificación ordinal y monotónica: [12]

- **Principio de dominancia**: a mayor valor en atributos de una instancia, mayor será el valor de la clase a la que se asigna dicha instancia. Usando el termino de "relación de dominancia" decimos que una instancia x domina a otra instancia x' cuando cada una de las variables de entrada de x (atributos de x) son mayores o iguales que cada uno de los de x', se denota $x \geq x'$ y por tanto x tendrá asignada una etiqueta de clase mayor que x'.
- **Función monótona**: una función es monótona si $x \geq x' \rightarrow h(x) \geq h(x')$

. Es decir, si x domina a x', la inferencia de clase de x ha de ser superior a la de x'.

Una vez definidos estos conceptos, podemos describir el sentido de la clasificación ordinal con restricciones monotónicas así como su diferencia con respecto a la clasificación ordinal simple:

- La **clasificación ordinal con restricciones monotónicas** maneja conocimiento subyacente del problema sobre clases ordenadas, atributos ordenados y una relación monotónica entre la evaluación de los atributos de una instancia y la asignación de esta a una clase.
- Si no hay relación de monotonía en la asociación de una clase a una instancia, pero las clases si poseen un orden, entonces se considera **clasificación ordinal** simplemente.

Con las restricciones de monotonía presentes se puede trabajar con una amplia variedad de funciones sin temor a que introduzcan más restricciones que la de monotonía: es posible hacer inferencia de la clase sobre todas las funciones monótonas.
La clasificación monotónica puede ser directa (más habitaciones, precio mayor de una casa), o inversa (más polución, precio menor de la casa).
Normalmente en problemas de clasificación monotónica reales, las restricciones monotónicas son consideradas en un subconjunto de características del dataset, no en todos los atributos.

Ejemplos de uso de monotonicidad: [9]

- Comparación de dos compañías donde una domina sobre la otra en términos de todos los indicadores financieros. Debido a esto, la compañía dominante ha de tener una evaluación final superior a la compañía dominada. Un uso de esto, es la predicción de la calificación crediticia usada por los bancos.
- House pricing: el precio de una casa será superior cuantas más habitaciones posea, mejor sea la calidad del aire acondicionado y menor sea la polución en el ambiente.

2.2.1. Restricciones monotónicas

La motivación del uso de restricciones monotónicas viene dada por los siguientes aspectos:

- El tamaño del espacio de la hipótesis es reducido, lo que facilita el proceso de aprendizaje.

- Otras métricas además de la precisión, como la consistencia con respecto a estas restricciones, pueden ser usadas por los expertos para aceptar o rechazar el modelo. Estas técnicas de evaluación de restricciones monotónicas las veremos más adelante con el fin de poder evaluar la consistencia de estas.

Las restricciones impuestas a continuación, son restricciones con respecto a la probabilidad de distribución en la generación de datos, además de imposiciones sobre la función de perdida bajo las cuales el clasificador óptimo de Bayes es monótono.

Dominancia estocástica

El principio de dominancia no siempre se aplica en la práctica de forma tan restrictiva, por lo que hemos de hablar en términos probabilísticos a la hora de referirnos a dichas restricciones.

Decimos entonces que, siendo 'k' una de las posibles clases a tomar en el dominio por una instancia 'x', y siendo 'y' la etiqueta asignada a dicha instancia 'x', si la restricción monótona nos dice que x≥x', entonces la dominancia estocástica nos dice que P(y≤k—x)≤P(y≤k—x'). Es decir, que la probabilidad de que el valor asignado (y) a la instancia dominante (x) sea mayor que cierto valor fijado de la clase (k), es mayor que la probabilidad de que el valor asignado (y) a la instancia dominada(x') sea mayor que ese mismo cierto valor de la clase fijado.

La relación de dominancia estocástica entre distribuciones se denota así:

$$x \succeq x' \Longrightarrow P(y|x) \succeq P(y|x')$$

Donde P(y|x) y P(y|x') denotan las distribuciones condicionales de la clase en x y x'.

Clasificador monótono de Bayes

En el problema de clasificación el objetivo es encontrar el clasificador más parecido al clasificador de Bayes, es decir, esta es nuestra función objetivo. Sabiendo esto, se convierte en requisito el hecho de que este también aplique las restricciones de monotonía que hemos enunciado.

Problema: aunque la distribución de probabilidad tiene restricciones monotónicas, el clasificador de Bayes no siempre las mantiene. Para solucionar este problema y mantener la monotonía en el clasificador de Bayes, han de imponerse las siguientes restricciones a la función de perdida (L):

- L(y,k+1)-L(y,k) ≥ L(y+1,k+1)-L(y+1,k)
  Esta característica de la función de pérdida es necesaria en la clasificación con restricciones monotónicas, si no no tendría sentido minimizar el riesgo dentro de la clase de las funciones monótonas.
  (Demostrado en [12])

- La siguiente definición de convexidad es necesaria también para mantener la restricción de monotonía en el clasificador de Bayes:

  - Siendo L(y,k) = c(y-k) (con c(0)=0)
  - La función c(k) es convexa si, para todo k entre -(k-1) y (k-1): c(k) ≤ (c(k-1)+c(k+1))/2
  - El clasificador de Bayes es monótono si y solo si c(k) (que es la V-shaped loss function) es convexa.

2.2.2. Métodos de clasificación no paramétricos

Los métodos no paramétricos son así llamados porque explotan la clase de todas las funciones monótonas. Estos métodos no hacen ninguna asunción más sobre el modelo que la de las restricciones monotónicas.

Aproximación Plug-In

Pretenden **estimar la distribución condicional de la clase**. Provienen de la clasificación isotónica (monótona creciente o decreciente)

Hemos de construir un método para estimar P(y|x), sabiendo que P(y|x) posee dos ventajas:

1. La distribución condicional permite la determinación de la predicción óptima para cualquier función de pérdida.

2. La distribución condicional mide la confianza de la predicción.

Problema de la clasificación binaria y la regresión isotónica.

En la aproximación plug-in se propone usar un vector de estimadores de densidad condicional p = (p1,....,pn), el cual es una regresión isotónica del vector de etiquetas y = (y1,...,yn). Es decir, p nos da la probabilidad de que x pertenezca a cada una de las clases existentes en y.

Dicho vector p es la solución del problema: $\sum_{i=1}^{n} (yi - pi)^2$ sujeto a las restricciones de monotonicidad (Xi≥Xj → pi≥pj). Por ello p minimiza el error cuadrático en el conjunto de los vectores monótonos p=(p1,..,pn) para cada x.

La elección de la función de error (función de perdida de error cuadrático) parece ser arbitraria. Puede verse que haciendo uso de otras funciones de perdida, se llega al mismo resultado.

La regresión isotónica es un problema de optimización cuadrática con restricciones lineales, por ello puede ser resuelta de forma eficiente con la mayoría de los resolutores de optimización de propósito general.

Problema multiclase.

Está basado en la regresión isotónica multiclase y, la idea es descomponer el problema de K-clases en varios problemas binarios y aplicar regresión isotónica a cada uno de los problemas. Está demostrado que la descomposición del problema de estimación de probabilidad para el caso de multiclase, siempre forma una adecuada distribución de probabilidad, es decir, que siempre son no negativos y la suma es igual a 1.

Aproximación directa

Consideramos la clasificación directa basada en la **minimización del riesgo empírico** dentro de la clase de todas las funciones monótonas.
Aunque este tipo de funciones no se pueden describir con un número finito de parámetros, la minimización del riesgo puede realizarse debido a que solo estamos interesados en valores de funciones monótonas en ciertos puntos, los incluidos en D (training set).

Una función monótona minimizando el riesgo empírico puede obtenerse resolviendo el siguiente problema de optimización:

- Minimizar: $\sum_{i=1}^{n} L(y_i, d_i)$.
- Teniendo en cuenta las restricciones de monotonía.
- Donde di son variables del problema (valores de la función monótona óptima en puntos de D)

El problema puede tener **otra interpretación** interesante: reetiquetar las instancias para hacer el dataset monótono de forma que las etiquetas de las instancias sean lo mas parecidas a las del conjunto original, donde esta similitud es medida en términos de la función de pérdida. Estas nuevas etiquetas serán los nuevos valores óptimos de las variables di. Este reetiquetado puede realizarse en el proceso de preprocesamiento y corresponde a la **corrección del error no paramétrico**.

Como el problema de clasificación no paramétrica se asimila al de la regresión isotónica (exceptuando que ahora se considera una salida discreta), será llamado ahora "clasificación isotónica" y su solución optima dŝerá llamada "clasificación óptima de y"

Capítulo 3

Contexto en algoritmos de árboles de decisión

3.1. Fundamentos de árboles de decisión

Los **árboles de decisión** [13] son un tipo de algoritmos de aprendizaje supervisado (tanto para clasificación como para regresión) utilizado en diversos ámbitos como la inteligencia artificial, las finanzas, el marketing, etc. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas en forma de ramificaciones de árboles, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Adentrándonos en los aspectos más técnicos de este tipo de modelos de predicción, cabe destacar que las variables de entrada y de salida pueden ser tanto categóricas como continuas y que divide el espacio de los predictores (variables independientes) en regiones distintas y no superpuestas, tal como veremos en la siguiente figura.

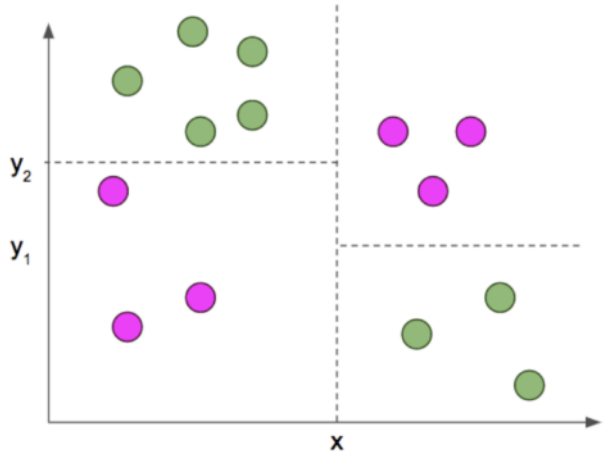

Figura 3.1: Ejemplo de árbol de decisión [1]

Estas divisiones se realizan creando sobre la población (el conjunto de datos) subconjuntos lo más homogéneos posible entre las muestras que componen un grupo y lo más heterogéneo posible entre los distintos subconjuntos.
Para la efectuación de esta separación, el algoritmo se basa en las variables de entrada más significativas, es decir, las que mejor separan las muestras.

A continuación podemos observar las diferentes partes de un árbol de decisión.

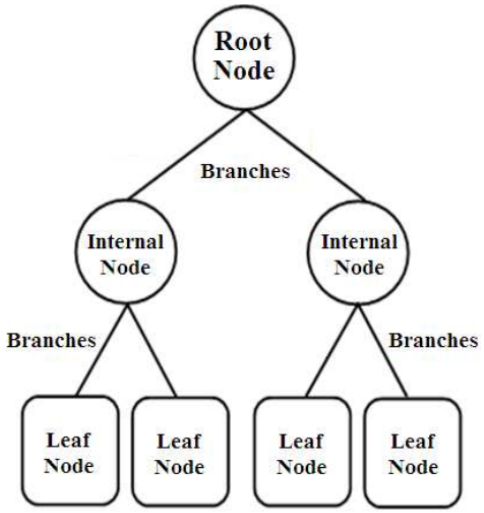

Figura 3.2: Partes de un árbol de decisión [16]

3.1.1. Terminología:

- **Nodo raíz**: Es el primero de los nodos del árbol y forma la población completa.
- **Ramificación**: Son las ramas que conectan todos los nodos del árbol por donde pasan las muestras para ser clasificadas.
- **Nodo de decisión**: Son aquellos donde las muestras se evalúan para decidir por qué rama continuar el camino hacia la solución.
- **Nodo terminal/hoja**: Estos son los nodos solución, una vez la muestra llega a este tipo de nodo, el proceso de evaluación de esta ya ha finalizado, por lo que habrá sido clasificada en alguno de los grupos categóricos existentes.
- **Poda**: Consiste en cortar u obviar una rama del árbol en la creación de un árbol, basándonos en cierta propiedad escogida, para evitar el recorrido del árbol completo y ahorrar de esta forma costos computacionales, así como para hacer frente al sobreajuste.
- **Rama/subárbol**: Es el conjunto de nodos y ramas completo que queda estrictamente por debajo de un nodo escogido del árbol total.
- **Nodos padre e hijo**: Dado un nodo del árbol, sus nodos hijo son todos aquellos que quedan conectados directamente a él únicamente en el nivel inferior siguiente. De esta forma, esos nodos hijo, comparten ese mismo padre.

3.1.2. ¿Regresión o clasificación?: similitudes y diferencias.

Similitudes

Ya sabemos, por ejemplo, que un árbol de decisión divide el espacio de los predictores en regiones no solapadas mediante el uso de los predictores más significativos.

Los árboles de decisión actúan bajo la llamada **separación binaria recursiva**, basada en un método greedy el cual decidirá en cada momento cuál será la mejor separación en el instante actual para encontrar el mejor árbol. El término 'binaria' hace alusión al tipo de división acaecido en cada nodo, es decir, que cada nodo divide en dos el espacio de los predictores. El término 'recursiva' se refiere a que el algoritmo realiza este proceso de forma reiterada hasta llegar a un criterio de parada predefinido.

Este proceso nos conduce a la generación de un árbol completo si no hacemos uso de criterios de parada, lo que nos lleva de forma directa al

problema del sobreajuste, obteniendo un modelo de una pésima calidad a la hora de evaluar nuevos datos. Por esto mismo es necesario difinir criterios de parada que realicen podas sobre el árbol para generar modelos lo suficientemente genéricos que eviten ese overfitting.

Diferencias

Las diferencias entre ambos modelos son bastante evidentes: para conjuntos de datos donde se utiliza una variable dependiente continua, utilizamos árboles de regresión, mientras que cuando la variable dependiente es categórica, usamos árboles de clasificación.
Dado esto, el **valor de los nodos hoja** no pueden ser calculados de la misma forma para ambas técnicas, por lo que, en **árboles de regresión**, utilizamos la **media** del valor de salida de las muestras que caen en dicho nodo hoja, mientras que en **árboles de clasificación** utilizamos la **moda** para asignar un valor de salida a nuevas muestras.

3.1.3. Ventajas e inconvenientes de los árboles de decisión

Ventajas

- Fáciles de comprender a la hora de interpretar los resultados.
- El tipo de dato utilizado no es una limitación.
- Es un método no paramétrico, es decir, en el que no es necesario hacer suposiciones sobre el espacio de distribución y la estructura del clasificador.
- Resulta útil a la hora de detectar la relevancia de los predictores aún habiendo una gran cantidad de estos.
- No son influidos por outliers ni valores perdidos (hasta cierto punto), por lo que requieren una menor limpieza de datos en comparación con otros métodos.

Inconvenientes

- Producen sobreajuste, por lo que hay que tener cuidado con ello mediante el uso de restricciones y la aplicación de poda.
- A la hora de trabajar con variables continuas, el árbol de decisión pierde información en el momento en el que categoriza dichas variables para la generación del árbol.

- No son del todo competentes con los mejores algoritmos de aprendizaje supervisado en cuanto a precisión en la predicción, es decir, no resultan ser tan efectivos ensambladores o SVM por ejemplo.

- Son sensibles al ruido en los datos, ya que este puede modificar de forma significativa la estructura del árbol.

3.1.4. Creación del árbol

Como ya sabemos, un árbol comienza desde un nodo raíz donde se encuentra clasificada toda la población y, conforme vamos profundizando por las ramas inferiores, vamos obteniendo subconjuntos cada vez más y más homogéneos con respecto a la variable de salida.

Para hacer posible esto necesitamos que nuestro modelo tome, por cada nodo, una decisión de separación de los datos basada en la ganancia de pureza (esa homogeneidad en los subconjuntos) al utilizar uno u otro predictor en cada uno de los nodos de decisión para crear esas particiones del espacio sucesivas.
Es decir, para cada nodo, se evalúa mediante unos medidores de pureza, cual es el predictor o característica del conjunto de datos en dicho instante que separa de mejor forma los datos que tenemos en ese momento con respecto a la variable de salida. Se escogerá en cada nivel, el predictor que mayor pureza ofrezca al árbol de decisión.

De esta forma vamos construyendo de manera progresiva ramas y más ramas del árbol haciendo uso de una técnica greedy de selección de característica a evaluar en cada nivel del árbol para la toma de decisión a la hora de generar nuevos nodos hijos.

¿Cómo medimos esa ganancia de homogeneidad?

Para los **árboles de regresión** sabemos que el objetivo de cada decisión del árbol en la creación de nuevas separaciones es minimizar la función RSS (**Residual Sum of Squares**) [5], una medida de error usada también en la regresión lineal. Es por ello que en cada nodo se escogerá una forma de particionar los datos mediante el uso de un predictor u otro, atendiendo a cuál minimiza en mayor medida dicha fórmula.

$$\mathrm{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2$$

Como el problema que nos concierne en este caso es el de **árboles de clasificación**, no entraremos en más detalles acerca de esta fórmula, ya que RSS no puede ser utilizado como criterio de separación binaria en este tipo de árboles.

Una aproximación natural a RSS es el '**ratio de error en la clasificación**', basado simplemente en la fracción de las observaciones de training en dicha región que no pertenecen a la clase más común de esta [4]. Se calcula de la siguiente forma:

$$E = 1 - \max(\hat{p}_{mk})$$

Donde $\widehat{P}$mk es la proporción $\widehat{P}$ de las observaciones de train de la región **m** que pertenecen a la clase **k**. Por tanto, el máximo de $\widehat{P}$mk hace alusión a la proporción máxima de elementos de train que siguen la moda en dicha partición.
Por desgracia esta medida de error no es lo suficientemente sensible conforme el árbol crece, por lo que es preferible el uso de otras medidas como el índice Gini y la Entropía.

Ya sabemos que buscamos nodos con una distribución de clase lo más homogénea posible. Con el fin de medir esa pureza en cada nodo, podremos utilizar los siguientes métodos ya mencionados:

- **Índice GINI**: nos indica como de pura es una región del espacio[5]. En este caso la pureza la definimos como la proporción de items de la región que pertenecen a una misma clase. Si la región contiene un índice alto de pureza, entonces el índice Gini será bajo (muy próximo a 0), siguiendo la siguiente fórmula:

$$G = \sum_{c=1}^{C} \hat{\pi}_{mc}(1 - \hat{\pi}_{mc})$$

  Donde $\widehat{\pi}$mc nos indica la proporción $\widehat{\pi}$ de items pertenecientes a la misma clase **c** en la región/nodo **m**.

- **Entropía (E-score)**: Se encarga también de la medida de homogeneidad de un nodo[5]. En este caso, los resultados obtenidos al aplicar la siguiente fórmula a cada nodo para observar el nivel de entropía, quedan más visibles con respecto a la medida Gini (es decir, se ve de forma más clara el nivel de homogeneidad de un nodo). Una entropía = 0

significa homogeneidad total, una entropía = 1 significa homogeneidad nula.

$$D = -\sum_{c=1}^{C} \hat{\pi}_{mc} \log \hat{\pi}_{mc}$$

¿Cómo evitamos el sobreajuste?

Una vez hemos escogido nuestra estrategia de creación del árbol, necesitamos indicarle al algoritmo cuándo ha de terminar de construirlo el uso de restricciones (prepruning) y su posterior poda (postpruning) para evitar de esta forma un sobreajuste a los datos.

Prepruning: establecimiento de parámetros

- Definir un número de observaciones mínimo sobre un nodo para que sea considerada una ramificación sobre él.
- Definir un número mínimo de observaciones sobre un nodo hoja.
- Establecer una profundidad vertical máxima para el árbol.
- Limitar el número máximo de nodos hoja.
- Parar si la expansión del nodo actual no mejora la medida de pureza utilizada actual.

Proceso de Postpruning:

1. Crear un árbol muy grande con o sin restricciones de prepruning.
2. Recorrer el árbol de abajo hacia arriba para ir cortando las hojas que nos dan ganancias negativas.

De esta forma podemos mantener ramas que, sin el proceso de postpruning podrían haber sido recortadas, pero que nos llevan a soluciones mejores que las que se ofrecen si este proceso por culpa del uso de la técnica greedy.

3.2. Hoeffding Tree y otros algoritmos de flujos de datos

Como ya sabemos, los algoritmos dedicados al data streaming han de seguir los siguientes requisitos:

- Procesar una muestra en cada momento y hacerlo tan solo una vez.
- Usar una cantidad de memoria limitada.
- Trabajar en un tiempo limitado.
- Estar listo para la predicción en cualquier momento.

Además, nuestro algoritmo ha de estar dotado de técnicas de detección de cambios en la distribución de los datos para evitar la disminución de la precisión en la predicción cuando esto suceda.

3.2.1. Hoeffding tree (VFDT)

Un árbol de Hoeffding es un algoritmo de inducción de árbol de decisión incremental capaz de aprender de flujos de datos masivos, suponiendo que la distribución que genera ejemplos es estacionaria, es decir, que no cambia con el tiempo. Los árboles Hoeffding explotan el hecho de que una pequeña muestra a menudo puede ser suficiente para elegir un atributo de división óptimo. Esta idea está respaldada matemáticamente por el Hoeffding bound, que cuantifica el número de observaciones (en nuestro caso, ejemplos) necesarias para estimar algunas estadísticas dentro de una precisión prescrita (en nuestro caso, la bondad de un atributo). [2]

Algunas de las técnicas de clasificación para data streaming tienen los siguientes **problemas**:

- Son altamente sensibles a la demanda de ejemplos.
- Carecen de alta eficiencia, siendo en algunos casos más lentos que un algoritmo batch.

Ante estos problemas se plantea **Hoeffding-tree** ya que:

- El aprendizaje de un Hoeffding-tree toma un tiempo constante en cada nuevo ejemplo, lo que lo hace adecuado para el aprendizaje de flujos de datos.
- Los árboles resultantes son similares a los creados con un batch learner convencional.

La cota de Hoeffding.

Hulten y Domingos presentan un método general para aprender de bases de datos grandes y arbitrarias. Este método consiste en derivar un límite superior para la pérdida del learner en función del número de ejemplos usados

en cada paso del algoritmo. De esta forma, se minimiza el número de ejemplos requeridos en cada paso del algoritmo, a la vez que se garantiza que el modelo obtenido no difiere de forma significante de aquel que se obtendría con todos los datos. Esta metodología de datos se ha aplicado de forma exitosa en k-means, clustering jerárquico de variables, árboles de decisión, etc.

Con el fin de cumplir con los requisitos establecidos al principio de este apartado para el tratamiento de flujos de datos, los autores proponen la cota Hoeffding para ser capaces de decidir la cantidad de instancias necesarias a evaluar para alcanzar un cierto nivel de confianza a partir del cual sabemos que no es necesario evaluar más ejemplos para seleccionar un atributo mediante el cual realizar la partición del árbol en el nodo actual.
Es decir, una vez alcanzada la cota de Hoeffding, el atributo que seleccionemos para el particionamiento del espacio de predictores, sera el mismo que seleccionaríamos si analizásemos una infinidad de ejemplos con el clasificador (evidentemente, con cierto nivel de confianza).

La idea básica consiste en usar un conjunto pequeño de ejemplos para seleccionar el test de divisón para colocar en un nodo del árbol de decisión. Si tras ver un conjunto de ejemplos, la diferencia en resultados entre ambos test de división no satisface un test estadístico (Hoeffding bound), entonces VFDT procede a examinar más ejemplos.

En VFDT se aprende un árbol de decisión de forma recursiva reemplazando hojas por nodos de decisión. Cada hoja almacena las estadísticas necesarias sobre los valores de los atributos. Dichas estadísticas necesarias son aquellas que se necesitan por una función de evaluación heurística que realiza el cálculo del resultado de los test de división basada en el valor de los atributos. Cuando hay un ejemplo disponible, atraviesa el árbol desde la raíz hasta una hoja evaluando el atributo requerido en cada nodo y siguiendo la rama correspondiente al valor del atributo en el ejemplo. Cuando el ejemplo llega a una hoja, la estadística de las hojas por las que ha pasado han sido actualizadas. Entonces cada condición basada en los valores de los atributos ha sido evaluada.

El nuevo nodo de decisión tendrá tantos descendientes como el número de posibles valores tenga el atributo escogido (por lo que el árbol no es necesariamente binario). Los nodos de decisión tan solo contienen la información sobre el test de división instalado en ellos.

Problema: VFDT no incluye soporte para el concept-drift por lo que, ante cambios en la distribución de los datos, los resultados del algoritmo pueden ser malos.

Cálculo de la cota:

Teniendo n variables independientes $r_1 ... r_n$ con un rango R y una media

$\bar{r}$, el Hoeffding bound afirma con una probabilidad 1-δ que la media real es al menos $\bar{r}$-ϵ donde:

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

Sabiendo que δ es la tolerancia al fallo a la hora de escoger un atributo en un nodo dado[14].
Una vez calculado el Hoeffding bound, se realizará la división si sucede una de las siguientes ocurrencias:

- La diferencia entre la ganancia ofrecida por el mejor atributo para la división y la ganancia ofrecida por el segundo mejor atributo, es mayor que el Hoeffding bound.

- El Hoeffding bound es menor que un parámetro de desempate (tie-breaking) establecido para cuando se dé el caso de que el Hoeffding bound es lo suficientemente pequeño pero los dos mejores atributos escogidos para la división son demasiado similares como para que se cumpla la condición anterior, lo que podría llevar a un estancamiento del árbol. El parámetro de desempate, según la literatura, suele establecerse a un valor de 0.05, que se suele alcanzar con unas 3.400 instancias analizadas.

3.2.2. Otros algoritmos de flujos de datos

Muchas bases de datos grandes presentan cambios en la distribución de los datos conforme avanza la generación de estos, es decir, posee un cambio de concepto en el tiempo, un **concept drift**. El hecho de que un algoritmo de flujos de datos no esté preparado para esos efectos cambiantes puede producir un empeoramiento de los resultados predictivos con el paso del tiempo, por lo que conviene tenerlo en cuenta a la hora de implementar esta clase de algoritmos.

Ante este problema, surgen técnicas como el uso de una **ventana deslizante** que tenga en cuenta los X ejemplos más recientes para asegurarnos de aprender siempre un modelo que tenga en cuenta el concepto actual se los datos, olvidando conceptos anteriores. Ante esta situación, ha de tenerse cuidado de escoger un valor adecuado de X, ya que ha de ser lo suficientemente grande como para tener un número suficiente de ejemplos con los que aprender el modelo y lo suficientemente pequeño como para abarcar un solo concepto de los datos.

CVFDT

Un algoritmo que utiliza este concepto de ventana deslizante es el llamado **Concept-adapting Very Fast Decision Tree (CVFDT)** que, evidentemente, contiene además las características del VFDT. A continuación la explicación de su funcionamiento[11]:

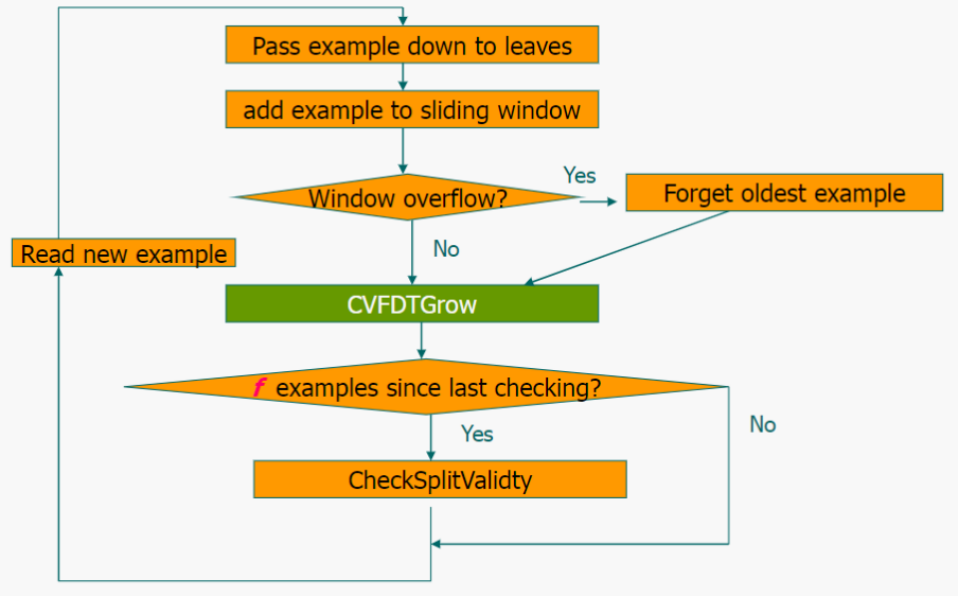

Figura 3.3: Ciclo CVFDT [6]

1. Entra un nuevo ejemplo al ciclo.
2. Se añade a la ventana deslizante.
3. Se comprueba si se excede la cantidad de ejemplos necesarios en dicho nodo para poder decidir el atributo apropiado para la división de los nodos hijos. En caso de excederse dicho tamaño de ventana, se elimina el ejemplo más antiguo.
4. CVFDTGrow: Se incrementan las estadísticas de cada uno de los nodos por los que pasa el ejemplo tanto en el árbol principal como en los subárboles. Si se ha alcanzado la cantidad suficiente de elementos analizados como para expandir un nodo hoja, escogemos el mejor de los atributos para expandir y se procede a ello.
5. CheckSplitValidity: Comprueba si comenzar un árbol alternativo o no basado en el atributo que mejor realiza la división en cada uno de los nodos. Se realiza una comprobación de las estadísticas de cada nodo observando si se produce un cambio de concepto con el nuevo ejemplo,

en cuyo caso es posible que el atributo escogido anteriormente en dicho nodo posea una ganancia menor que otro, por lo que el algoritmo creará un subárbol alternativo cuya raíz será ese atributo que tenga mejor gain que el actual en dicho nodo.

Cada uno de los nodos de decisión ha de tener las estadísticas suficientes para poder olvidar información en caso de que se produzca un cambio de concepto. Para ello, cada uno de los nodos posee un ID monotónicamente incremental que permite al algoritmo validar las decisiones previamente tomadas y realizar el olvido de información inútil en los cambios de concepto. Cuando un ejemplo antiguo es eliminado de la ventana, se recorre el árbol para decrementar las estadísticas por cada uno de los ID de los nodos por los que ha pasado.

HATT

Otro algoritmo que implementa una solución más eficiente a la presentada por Hoeffding tree es el **Hoeffding Anytime Tree** [14], el cual asegura conseguir mejores resultados explotando la información conforme le llega al modelo en lugar de esperar a traspasar el Hoeffding bound y, realizando correcciones sobre estas decisiones de elección de atributo para la división siempre que sea necesario. Por tanto, en cuanto se detecte una división útil en cualquiera de las hojas del árbol, se realizará de forma inmediata y esta, será reemplazada tan pronto como otra alternativa mejor sea identificada.

Además de esto, algunos experimentos muestran que el algoritmo, aún no habiendo sido tratado para ello, muestra algunas características de tratamiento de concept drift.

Option trees

Con el objetivo de introducir a Hoeffding tree mayor estabilidad y romper la barrera que posee VFDT para mirar hacia adelante, surje Option trees[15].
Consiste en una estructura que representa a múltiples árboles en lugar de a uno solo como hace VFDT. Una instancia nueva puede bajar por varios caminos del árbol contribuyendo de diferentes maneras en diferentes opciones. La clase de un ejemplo de test se determina por un comité(mayoría de voto o pesos) hecho por las predicciones de todas los nodos hoja alcanzados. El concepto es crear múltiples opciones pero de la misma forma que lo hace Hoeffding tree.
Esta nueva representación de un árbol difiere únicamente de la representación hecha para VFDT en que contienen unos nodos llamados **nodos opción**

que se encargarán de evaluar a los ejemplos que pasen por ellos mediante múltiples tests para determinar por qué ramas dejarle pasar.
En este nuevo método, un ejemplo que entra al árbol puede influir en varios nodos hoja distinto, mientras que en el Hoeffding tree, un solo ejemplo tan solo puede influir en un solo nodo hoja.

3.3. Árboles de decisión monotónicos

Conociendo las restricciones monotónicas descritas en el anterior capítulo, hemos de ser capaces de extrapolar dichos conocimientos a los árboles de decisión para poder resolver problemas de este estilo.

Recordamos que, siendo X e Y valores de atributos y C_x y C_y las clases asignadas a X e Y respectivamente, se cumple una **relación de monotonía** entre el par atributo-clase (X,C_x) y el par (Y,C_y) si y solo si (X,C_x) domina a (Y,C_y), viceversa o son exactamente iguales (tanto en valores de sus atributos como en clase asignada).

Hacemos uso de **este conocimiento sobre un árbol de decisión** de la siguiente manera:
Siendo (P,C_p) y (Q,C_Q) dos caminos distintos del mismo árbol de decisión (donde P y Q son atributos y C_p y C_q son nodos respuesta), estos son monotónicos entre sí, si cumplen las mismas reglas de monotonía descritas justo en el párrafo anterior (relación de dominancia).[8]

3.3.1. Problema:

Un conjunto de datos donde todos sus ejemplos guardan una relación de monotonía entre sí, no garantiza que genere un árbol de decisión monotónico a través de algoritmos teóricos TDIDT(top-down induction decision tree) que usan la entropía como selector de atributos.

Crear un árbol de decisión que cumpla las restricciones de monotonía estudiadas y que al mismo tiempo vele por la minimización de error (la precisión) no es tarea sencilla.

3.3.2. ¿Cómo creamos un árbol de decisión monotónico?

Aún no siendo una tarea simple, existen **métodos de creación de estos árboles** como los que describimos a continuación.

Método basado en matriz

Uno de estos métodos es el **basado en matriz**. Teniendo ya construido un árbol con k ramas, construimos una matriz simétrica M de tamaño k x k donde el valor m_{ij} es un 1 en caso de que la rama de la fila i no guarde una relación de monotonía con la rama de la columna j y un 0 en caso contrario. Cada columna (y fila) está asociada con un contador que hace referencia a la suma de los unos que contiene, de esta forma sabemos con qué cantidad de ramas no guarda la relación de monotonía. Comenzamos eliminando (podando) aquellas ramas del árbol que tienen un contador mayor de no-monotonía y actualizando los contadores del resto de ramas hasta llegar a obtener una matriz llena de ceros o hasta llegar a una matriz 1 x 1, en cuyo caso M sería una matriz de no-monotonicidad.

Método aleatorio

Otro método más rápido a la hora de ejecutar es tomar inicialmente de forma aleatoria una rama del árbol y declararla como monotónica. A partir de aquí ir tomando siempre de forma aleatoria cada una de las demás ramas del árbol, compararlas con las ramas ya declaradas monotónicas y, o bien descartarlas (en caso de no guardar relación de monotonía con las ya declaradas monotónicas) o bien introducirlas en el conjunto de ramas declaradas monotónicas.

Método de evaluación de monotonicidad y precisión para expansión

Describimos en este apartado una métrica para que tenga en cuenta tanto el error como las restricciones monotónicas a la hora de decidir si expandir un nodo del árbol de decisión, lo que lo convierte en el tercero de los métodos que describiremos para la creación de árboles de decisión montónicos[8].

Primeramente definimos una medida de no-monotonicidad para los árboles de decisión:
El **índice de no-monotonicidad** nos dice el ratio entre el número real de pares de ramas no monotónicas de un árbol de decisión y el número máximo de pares que podrían no haber sido monotónicos con respecto a otras en el mismo árbol.

Para conseguir este índice hacemos uso de la mátriz M creada en el apartado anterior y denotamos W como la suma de todas las entradas de la matriz M, es decir:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}$$

Sabemos que, como mucho, (k^2 - k) entradas de M pueden ser etiquetadas como no monotónicas, por tanto el índice de no-monotonicidad queda así:

$$I_{a_1,a_2,\ldots,a_v} = \frac{W_{a_1,a_2,\ldots,a_v}}{k^2_{a_1,a_2,\ldots,a_v} - k_{a_1,a_2,\ldots,a_v}}$$

Previo cálculo del índice que nos dirá como de bueno es un árbol tanto en precisión como en aguardar las restricciones monotónicas, calculamos el **order-ambiguity-score**, que se define en términos del índice previamente calculado tal como sigue:

$$A_{a_1,a_2,\ldots,a_v} = \begin{cases} 0 & \text{if } I_{a_1,a_2,\ldots,a_v} = 0 \\ -(\log_2 I_{a_1,a_2,\ldots,a_v})^{-1} & \text{otherwise} \end{cases}$$

Finalmente, añadimos esta medida de no-monotonicidad calculada a nuestra medida de precisión E-score de la siguiente manera:

$$T_{a_1,a_2,\ldots,a_v} = E_{a_1,a_2,\ldots,a_v} + A_{a_1,a_2,\ldots,a_v}$$

Una vez tenemos creada la métrica, ya sabemos que, a menor valor, mejor será el resultado, ya que significará que poseemos un menor fallo en la predicción y una mayor conservación de las relaciones de montonía dentro del árbol.

Esta forma de evaluar un árbol de decisión no implica que las consideraciones de monotonía deban dominar la construcción del árbol necesariamente. Podemos decidir cuánta importancia darle al proceso de construcción de un árbol montónico multiplicando el order-ambiguity-score por un factor para darle menor importancia al hecho de que el árbol aguarde las restricciones monotónicas con respecto al de que consiga una buena precisión en la predicción. [8]

Capítulo 4

Propuesta

4.1. Introducción

Una vez conocemos el funcionamiento, características, algunas formas de implementación y los problemas a los que se enfrentan cada una de las técnicas descritas hasta ahora (árboles de decisión, data streaming y la aplicación de restricciones monotónicas), procedemos a exponer la propuesta hacia la cual se encauza el propósito de este documento.

Hemos comprobado que existen métodos eficientes de adquisición de conocimiento y predicción sobre flujos de datos utilizando árboles de decisión, tales como Option trees y Hoeffding Anytime Tree (HATT), ambos basados en Hoeffding Tree, o incluso su versión mejorada para la detección del cambio de concepto en el tiempo (CVFDT).

Así mismo, hemos observado también que existen ejemplos de conjuntos de datos que poseen características dignas de ser tenidas en cuenta a la hora de crear un modelo de aprendizaje de conocimiento, ya que aportan una base de información subyacente a los datos que ofrece al algoritmo la capacidad de crear una estructura de datos más eficiente para la resolución de cualquier problema que se enfrente a datos de este estilo. Evidentemente hablamos de los conjuntos de datos que aguardan en sí una relación de monotonía entre los valores de los atributos de sus instancias y las clases asignadas a estas.

Tal como se muestra en [8], existen multitud de problemas en la vida diaria que requieren ser tratados como problemas de clasificación monotónica con el fin de ser correctamente resueltos. Este es el caso de aquella universidad que no quiera admitir a un solicitante con ciertas notas y rechazar a otro con notas iguales o más altas por no haber tenido en cuenta la monotonicidad

del asunto, o el caso de una compañía de seguros de vida que no desea que sus decisiones dependan de un árbol de decisión que no tenga en cuenta que un solicitante anciano y poco saludable ha de cotizar una tasa de prima más alta que un solicitante joven y saludable. Otros problemas pueden ser los de credit scoring, elección de consumidor, selección de escuela y transporte, etc.

En el siguiente ejemplo veremos las consecuencias que puede tener el hecho de crear un árbol de clasificación no-monotónico en un problema de credit scoring simple. Los atributos utilizados en cada árbol son los ingresos(income) + los activos(assets) y los activos a solas respectivamente.

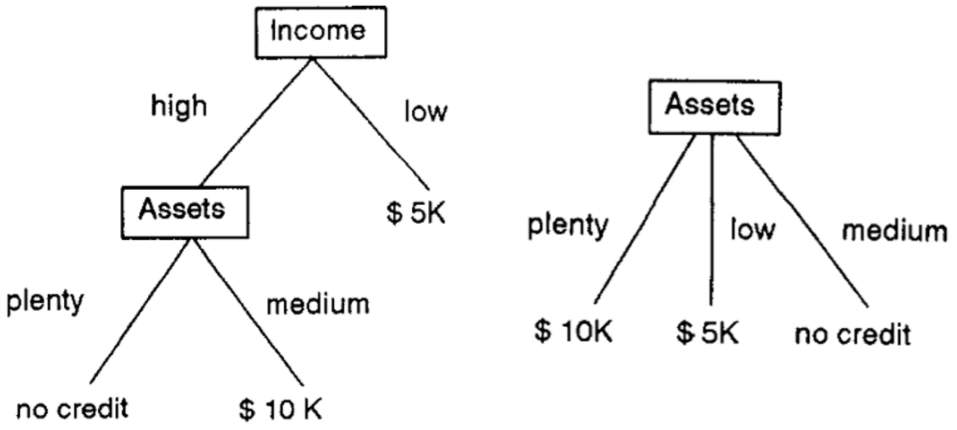

Figura 4.1: Árboles no monotónicos[8]

Como vemos, ambos árboles, al no poseer restricciones de monotonicidad, carecen también de sentido, ya que a un cliente con pocos activos se le autoriza una linea de crédito de cinco mil dólares, mientras que a un cliente con más activos no se le ofrece crédito alguno, por ejemplo.

El árbol no-monotónico de la derecha en la figura anterior, es el resultado surgido del siguiente conjunto de datos que **sí** guarda relación de monotonía entre los valores de los atributos y las clases asignadas. Tal como vemos, el hecho de que los datos sí sean monotónicos no asegura que el árbol creado a partir de él lo sea.
Los datos utilizados para ello son los siguientes:

Example	Income	Assets	Credit history	Class
#1	high	plenty	good	$ 10K
#2	high	low	bad	$ 5K
#3	low	medium	bad	no credit

Figura 4.2: Conjunto de datos monotónico[8]

4.2. Propuesta y resultados esperados

Visto todo esto, parece lógico crear una adaptación de los algoritmos de árboles de decisión para data streaming que además posean información sobre restricciones de monotonía para conseguir resultados superiores en cuanto al análisis del dominio del problema al que se enfrentan los algoritmos que trabajan con este tipo de conjuntos de datos.

La propuesta no pretende conseguir resultados mejores en cuanto al accuracy obtenido en la predicción de resultados si no que, pretende que las estructuras creadas por los árboles de decisión para flujos de datos sean más fieles a la realidad que subyace bajo data sets con cierto nivel de monotonía en su relación atributos-clase, es decir, que se creen estructuras con sentido lógico, en lugar de árboles de decisión que obtengan buenos resultados en precisión pero malos en el concepto del dominio real que maneja el problema, lo que nos lleva a conseguir árboles con poca o nula interpretabilidad a la hora de la verdad.

Dicho esto, podría darse el caso en el que el algoritmo básico comparativo sin restricciones de monotonía aplicadas alcance un accuracy superior al que puedan alcanzar las adaptaciones realizadas las cuales si posean información sobre dichas relaciones monotónicas, por lo que, aunque también mediremos el nivel de acierto evidentemente, el objetivo principal de las mejoras a realizar con los novedosos algoritmos es conseguir un nivel mayor de monotonicidad y, por tanto, de lógica en el análisis del problema por parte de nuestros algoritmos, lo que les proporcionará un nivel de conocimiento superior del problema al que se están enfrentando.

Continuando con el ejemplo del credit scoring, a continuación veremos el árbol generado por un algoritmo que sí hace uso de información monotónica a la hora de su construcción. Se ve de forma rápida en esta figura que la lógica está presente en la resolución del problema.

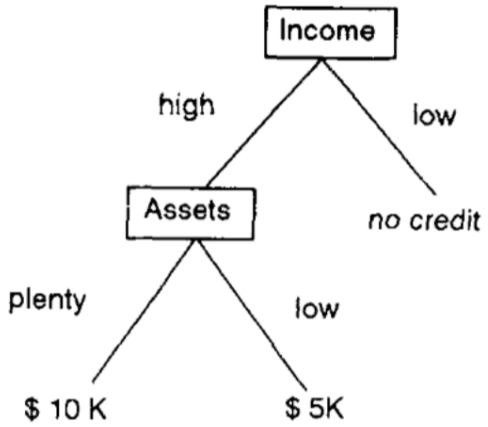

Figura 4.3: Árbol creado con restricciones monotónicas sobre un conjunto de datos monotónicos [8]

A continuación podemos ver el pseudocódigo de la propuesta, el mismo que se ha utilizado para la implementación de los experimentos a nivel de programación:

Algorithm 1 Árboles de clasificación monotónica sobre flujos de datos. Método de matriz de colisiones.

```
1: procedure ENTRENAMIENTO
2:     while lleganDatosEntrenamiento do
3:         modelo ← entrenarHoeffdingTree(nuevaInstancia)
4:         if IteracionesParaPoda then
5:             -Crear matriz de colisiones entre ramas
6:             -Podar ramas con mayor indice de colisión
7: procedure VALIDACIÓN
8:     while lleganDatosTest do
9:         resultados ← validar(nuevaInstancia, modelo)
10:    MAE ← resultados.accuracy
11:    NMI ← resultados.indiceNoMonotonicidad
```

Algorithm 2 Árboles de clasificación monotónica sobre flujos de datos. Método de evaluación a la hora de expandir el árbol.

```
    procedure ENTRENARHOEFFDINGTREE(NUEVAINSTANCIA)
2:      Bajar instancia hasta nodo hoja
        Calcular cota Hoeffding
4:      Calcular mejores atributos para la división del nodo en varios hijos
        Calcular NMI para dicho nodo hoja
6:      if SuperadaCotaHoeffding & obedeceMonotonicidad then
            Expandir nodo en varios hijos
8:  procedure ENTRENAMIENTO
        while lleganDatosEntrenamiento do
10:         modelo ← entrenarHoeffdingTree(nuevaInstancia)
    procedure VALIDACIÓN
12:     while lleganDatosTest do
            resultados ← validar(nuevoDato, modelo)
14:     MAE ← resultados.accuracy
        NMI ← resultados.indiceNoMonotonicidad
```

4.3. Algoritmos a comparar

Como **algoritmo de comparación** utilizaremos aquél del cual la literatura ha hecho más uso para la realización de estudios comparativos y mejoras de otros aspectos, es decir: **Hoeffding Tree**, la base de la que parten gran cantidad de algoritmos de árboles de decisión para el tratamiento de data streaming.

Los **algoritmos que pretenden hacerle frente** en cuanto a los aspectos ya mencionados serán:

- Adaptación de Hoeffding Tree con restricciones monotónicas mediante el uso del **método basado en matriz**, el cual sabemos que primeramente crea el árbol de decisión para, posteriormente, aplicarle sucesivas podas hasta conseguir un árbol monotónico en caso de ser posible.

- Adaptación de Hoeffding Tree con restricciones monotónicas mediante el uso del **método de evaluación de monotonicidad de ramas** que, como bien hemos explicado anteriormente, en la propia construcción del árbol, a la hora de decidir si expandir una rama o no, se basará tanto en la medida usada para comprobar el nivel de accuracy como en la medida de monotonicidad de ramas(índice de no-monotonicidad).

- Adaptación de Hoeffding Tree con restricciones monotónicas mediante el uso de **ambos métodos al mismo tiempo**.

Capítulo 5

Software desarrollado y uso

El software para la realización de los experimentos ha sido creado mediante el IDE de programación Netbeans con Java como lenguaje de programación.

Los datos empleados son estáticos, por lo que ha sido necesario adaptar el código fuente a un entorno de flujo de datos para poder realizar los experimentos.

Como ya se ha comentado, el algoritmo de partida de los experimentos es el Hoeffding Tree, el cual se basa en una adaptación de árboles de decisión para flujos de datos. Este código ha sido tomado de la API de MOA.

MOA (Massive Online Analisys) [3] es un entorno de trabajo open-source relacionado con el proyecto WEKA (Waikato Environment for Knowledge Analysis) para el tratamiento o minería de flujos de datos de evolución masiva conteniendo una gran colección de algoritmos de machine learning como son: clasificación, regresión, clustering, detección de outliers, detección de concept drift y sistemas de recomendación, así como herramientas de evaluación.

Los **algoritmos implementados** son los descritos en la propuesta:

- Algoritmo 1: Árboles de clasificación monotónica sobre flujos de datos. Método de matriz de colisiones.
- Algoritmo 2: Árboles de clasificación monotónica sobre flujos de datos. Método de evaluación a la hora de expandir el árbol.

Para la obtención de resultados y la posible comparativa de estos entre los distintos modelos, para ambos algoritmos ha sido necesario modificar algunos de los parámetros de Hoeffding Tree:

- **Desempate(tie-breaking)**: este parámetro ayuda, tal como su pro-

pio nombre indica, a evitar estancamientos en las expansiones de los árboles acaecidos por la similitud entre la ganancia ofrecida por dos atributos a la hora de decidirse por alguno para expandir. Si son muy similares, la diferencia entre ellos sera pequeña, por tanto será imposible que supere la cota Hoeffding, lo cual es necesario para poder expandir. Gracias a su modificación, he podido obtener árboles más extensos, ya que sin tocarlo, el pequeño volumen de datos manejado no lo permitía.

- **Grace period option**: con él indicamos la cantidad de instancias necesarias para la generación de nodos nuevos en el árbol.

Dichos han sido creados adaptando el código y resultados de Hoeffding Tree a las necesidades de cada uno y, los resultados de estos han sido sometidos a **validación cruzada de 10 iteraciones** para obtener resultados más fieles al comportamiento de cada algoritmo. Para obtener los resultados de cada iteración de la validación cruzada he creado mis propias **funciones de evaluación** de la medida de **accuracy (MAE en este caso)** y del **índice de no-monotonicidad (NMI).**

Para la obtención de la medida **MAE**, como ya hemos visto, simplemente divido la suma de errores de cada iteración entre la cantidad de elementos en el conjuntos de datos.

Para obtener el NMI hago uso de los datos predichos por el algoritmo de la siguiente forma:

1. Detecto las colisiones entre las distintas instancias (es decir, compruebo con cuantas instancias no aguarda monotonía respecto a atributos y clase cada una de las instancias del conjunto de datos).
2. Realizo la suma de dichas colisiones.
3. Dividio entre la cantidad máxima de colisiones que podrían ocurrir en dicho conjunto de datos.

El **proceso de aprendizaje del modelo y validación del mismo** para cada iteración de la validación cruzada es el siguiente:

1. Entrenamiento del modelo con el algoritmo escogido de los tres implementados haciendo uso de los datos de entrenamiento a modo de flujo de datos.
2. Generación de predicciones con los datos de test.
3. Validación de las predicciones en cuanto a accuracy (MAE) y monotonicidad (NMI).

Estos resultados obtenidos son los que se expondrán más adelante en la comparativa de algoritmos.

Capítulo 6

Experimentos

Tras el análisis introductorio al contexto en el que nos situamos para hacer conocer al lector las técnicas que hay detrás del problema que queremos resolver y después de exponer y analizar la propuesta de trabajo a la que nos enfrentamos, entramos de lleno en la exposición del marco de trabajo en el que nos situaremos con los algoritmos que vamos a manejar, así como la muestra de resultados obtenidos en cuanto a comparativas realizadas y, por último, su posterior análisis, con el fin de concretar el éxito o fracaso de la teoría expuesta en capítulos anteriores.

6.1. Marco de trabajo

En esta sección, tal como hemos comentado, situaremos el marco de trabajo de nuestro experimento, formado por los conjuntos de datos que trataremos para este, junto con las medidas de precisión y monotonía que usaremos con los algoritmos con el fin de poder ser comparados por igual.

6.1.1. Conjuntos de datos

Primeramente comentaremos los conjuntos de datos a tratar con el fin de entender a que se enfrentarán nuestros algoritmos. Antes que nada cabe destacar que los data sets que vamos a utilizar son estáticos, es decir, no son datos que nos llegan en flujo (que es uno de los aspectos que queremos tratar en este documento), por tanto a la hora de realizar los experimentos habremos de simular la entrada de datos por un flujo en lugar de leerlos de forma directa. Los datos han sido sacados de [7]

Data set	Instancias	Atributos	Numéricos	Nominales	Clases	NMI
ERA	1000	4	4	0	9	0.016
ESL	488	4	4	0	9	0.004
LEV	1000	4	4	0	5	0.006
SWD	1000	10	10	0	4	0.009

Cuadro 6.1: Conjuntos de datos a utilizar en el experimento

Descripción de los conjuntos de datos

- **ERA**: recopilación de datos tomados durante un experimento académico de toma de decisiones académicas con el objetivo de determinar las cualidades más importantes de los candidatos para ciertos tipos de trabajo.
- **ESL**: perfiles de aplicantes para ciertos trabajos en la industria. Psicólogos expertos determinaron los valores de los atributos de los datos basándose en resultados de tests psicométricos y entrevistas realizadas a los candidatos.
- **LEV**: estos datos hacen referencia a ejemplos de evaluaciones anónimas de profesores realizadas al final de cursos MBA (Master of Business Administration). Antes de recibir las notas finales, se le pidió a los estudiantes que puntuaran a sus profesores en concordancia a cuatro atributos como las habilidades orales y la contribución a su conocimiento profesional o general. La salida es una evaluación total del rendimiento del profesor.
- **SWD**: este conjunto de datos contiene evaluaciones reales de trabajadores sociales cualificados midiendo el riesgo de un grupo de niños al permanecer en casa con sus familias. Esta evaluación de riesgos se presenta a menudo en cortes judiciales para ayudar a decidir que le interesa más a un niño presuntamente maltratado o descuidado.

6.1.2. Medidas

Primeramente abordaremos las medidas destinadas a valorar la precisión y, posteriormente, aquellas cuyo objetivo es medir la monotonicidad [9].

Medidas de precisión

Dentro de las medidas de precisión para clasificación podemos encontrarlas de dos tipos: para clasificación binaria y para clasificación multiclase.

Clasificación binaria

Previa exposición de algunos de las medidas de clasificación binaria necesitamos tener en cuenta algunos términos con el fin de poder entenderlas:

- **True Positives (TP)**: cantidad de instancias con predicción positiva que están correctamente clasificadas.
- **False Positives (FP)**: cantidad de instancias con predicción positiva que no están correctamente clasificadas.
- **True Negative (TN)**: cantidad de instancias con predicción negativa que están correctamente clasificadas.
- **False Negative (FN)**: cantidad de instancias con predicción negativa que no están correctamente clasificadas.

Una vez conocemos esta terminología, vemos algunas de las medidas para clasificación binaria:

- **Accuracy**: Representa la habilidad predictiva de acuerdo con la proporción de los datos clasificados de forma correcta.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Error rate**: Caso opuesto al accuracy, evaluando la proporción de los datos evaluados clasificados de forma incorrecta.

$$\text{Error rate} = \frac{FP + FN}{TP + FP + TN + FN}$$

- **Recall/sensitivity**: Medida de la proporción de TP que están correctamente clasificados.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Positive predictive value(PPV)/Precisión**: Proporción de las instancias del test que tienen una salida positiva y que además están bien clasificados. representa la probabilidad de que una prueba positiva refleje la condición subyacente que se está probando.

$$\text{PPV} = \frac{TP}{TP + FP}$$

Clasificación multiclase

Procedemos ahora a exponer algunas medidas de predicción aplicadas a los problemas de clasificación multiclase. Para ambas medidas hemos de tener en cuenta lo siguiente: n corresponde a la cantidad de observaciones en el conjunto de datos evaluados, y'_i es la clase predicha para una instancia i e y_i es la etiqueta de clase real (ambos representados como valores enteros basados en su posición en la escala ordinal).

- **Mean Squared Error (MSE)**: Mide la media de los cuadrados de los errores. Al usar los cuadrados de y_i e y'_i se pondera con un peso mayor a los errores más grandes.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2$$

- **Mean Absolute Error (MAE)**: Mide cómo de cerca se encuentran las predicciones de los valores reales de salida. MAE es una medida lineal, lo que significa que los errores son tratados con el mismo peso en la media.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y'_i - y_i|$$

Dados los conjuntos de datos de los que disponemos y el uso que la literatura recomienda hacer de las medidas de precisión sobre este tipo de conjuntos de datos, utilizaremos la medida **Mean Absolute Error (MAE)** para la evaluación de la precisión de nuestros algoritmos, ya que los conjuntos que manejaremos son de clasificación multiclase.

Medidas de monotonicidad

Nuestro propósito con esta medida es evaluar la tasa de monotonicidad provista tanto por las predicciones obtenidas y el conjunto de datos original, como de la construcción del modelo.

Para ello tenemos en cuenta las siguientes consideraciones:

- x es un ejemplo del conjunto de datos D.
- NClash(x) es la cantidad de ejemplos de D que no cumplen las restricciones de monotonicidad con respecto a x.

- n es el número de instancias en D.
- NMonot(x) es el número de ejemplos de D que sí que cumplen las restricciones de monotonicidad con respecto a x.

El **Índice de no-monotonicidad** se define como el número de choques (Nclash) dividido por el número de pares de ejemplos en el conjunto de datos:

$$NMI = \frac{1}{n(n-1)} \sum_{\mathbf{x} \in D} NClash(\mathbf{x})$$

Y esta será la medida que utilicemos finalmente para la evaluación de la monotonicidad en nuestros algoritmos.

6.2. Resultados

A continuación muestro las condiciones bajo las que se han ejecutado todas las tablas de resultados que mostraré a continuación. Estas condiciones son los parámetros de Hoeffding Tree ajustados para obtener un árboles lo suficientemente representativos de los datos.

- Parámetro de desempate: 1
- Parámetro grace period:
  - ERA = 20
  - ESL = 45
  - LEV = 20
  - SWD = 20

Como era de esperar, el conjunto de datos ESL, al ser bastante más pequeño que los demás, necesita un aumento del parámetro grace period.

Para el segundo algoritmo, el que trata los índices de monotonicidad de los nodos para decidir si hacer una expansión de este o no, posee además un parámetro de tolerancia de no-monotonicidad para evitar intentar conseguir árboles muy monótonos a costa de la pérdida de precisión. El parámetro es un simple factor del NMI calculado en cada nodo con respecto a las demás ramas. Los ajustes de este parámetro para los experimentos han sido calculados de forma empírica hasta obtener unos resultados óptimos. Estos son:

- ERA = 0.5
- ESL = 0.45
- LEV = 1
- SWD = 0.45

HF	ERA	ESL	LEV	SWD
MAE	1.673	0.580	0.612	0.625
NMI	0.4060	0.3856	0.3473	0.3510

Cuadro 6.2: Resultados de accuracy y monotonicidad para los 4 data sets descritos mediante el uso de Hoeffding Tree algorithm.

HF + Poda	ERA	ESL	LEV	SWD
MAE	1.88	0.576	0.636	0.66
NMI	0.3892	0.3849	0.3382	0.342

Cuadro 6.3: Resultados de accuracy y monotonicidad para los 4 data sets descritos mediante el uso de Hoeffding Tree algorithm junto con poda.

HF + Decisión	ERA	ESL	LEV	SWD
MAE	1.65	0.642	0.562	0.62
NMI	0.3957	0.3791	0.3369	0.3458

Cuadro 6.4: Resultados de accuracy y monotonicidad para los 4 data sets descritos mediante el uso de Hoeffding Tree algorithm junto con el algoritmo de toma de decisión a la hora de expandir.

HF + Poda + Decisión	ERA	ESL	LEV	SWD
MAE	1.67	0.64	0.664	0.629
NMI	0.3815	0.3790	0.3451	0.3447

Cuadro 6.5: Resultados de accuracy y monotonicidad para los 4 data sets descritos mediante el uso de Hoeffding Tree algorithm junto con ambos algoritmos creados.

6.3. Análisis

A la vista de los resultados podemos observar que, aunque no en gran medida, la adición de restricciones monotónicas a los algoritmos de flujos de datos con árboles de decisión resulta ser exitosa, ya que el índice de no-monotonicidad desciende al aplicar cualquiera de los algoritmos creados, ya sea el de poda, el de toma de decisión a la hora de expandir los nodos o ambos al mismo tiempo.

Hemos de tener en cuenta también que los conjuntos de datos que estamos tratando en los experimentos son muy pequeños, el entorno de flujos de datos real ofrece cantidades masivas de datos que permiten a los algoritmos un aprendizaje más fiel del problema, quizá este sea el motivo de que el índice de monotonicidad descienda tan poco entre el algoritmo sin las restricciones y los que sí las tienen. Al no tener tantos datos hemos debido de modificar los parámetros de Hoeffding Tree para poder crear árboles lo suficientemente grandes que nos sirvan para poder probar nuestros algoritmos.

Por contra, podemos observar cómo en algunas ocasiones, un descenso de la monotonicidad viene acarreado de sanciones en la precisión del algoritmo, por lo que el índice MAE puede verse ligeramente mayor a la hora de aplicar estas restricciones monotónicas. Como dijimos capítulos atrás, el objetivo del presente documento era la disminución de la monotonicidad de los resultados predichos por nuestros algoritmos. Es probable que el algoritmo base, sin las restricciones de monotonía, alcance una precisión más alta pero que la interpretabilidad de los resultados predichos sea inferior a la de los resultados obtenidos por los algoritmos que sí aplican dichas restricciones.

En definitiva, teniendo en cuenta que el factor que hace que los resultados no sean tan significativos es el volumen de los conjuntos de datos tratado, podemos concluir que los resultados han sido exitosos ya que, aunque hemos obtenido mayor error en la predicción, la interpretabilidad de los resultados es mayor debido a la bajada de la no-monotonicidad en las predicciones.

Capítulo 7

Conclusiones y trabajo futuro

Como conclusión del estudio completo realizado sobre introducción de restricciones monotónicas a los algoritmos de clasificación de flujos de datos con árboles de decisión para el tratamiento de conjuntos de datos con este tipo de relaciones entre los atributos y la variable de salida, hemos podido comprobar que, en efecto, resulta útil para obtener modelos más fieles al problema que se trata con dicho tipo de algoritmos, pero tampoco han parecido tener un éxito demasiado significante si hablamos de números, al menos con los conjuntos de datos empleados.

Es por aquí por donde merece la pena, incluso es necesario, continuar este objeto de estudio, es decir, tratar de ver los resultados de estos algoritmos con flujos de datos reales, donde el volumen masivo nos permitirá comprobar su real efectividad y nos dirá si es necesaria la aplicación de esta novedosa técnicas sobre problemas reales. Es muy posible que al utilizar flujos de datos reales los resultados sean más significativos y consigamos una brecha mayor en cuanto a índice de monotonicidad entre el algoritmo base (sin uso de restricciones) y los demás creados con este fin.

Bibliografía

[1] Decision tree basics. https://bookdown.org/content/2031/arboles-de-decision-parte-i.html#que-son-los-arboles-de-decision.

[2] Definición hoeffding tree. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/HoeffdingTree.html.

[3] Definición moa. https://moa.cms.waikato.ac.nz/.

[4] Fórmula cer. https://towardsdatascience.com/everything-you-need-to-know-about-decision-trees-8fcd68ecaa71.

[5] Fórmula rss, gini y entropía. https://www.quantstart.com/articles/Beginners-Guide-to-Decision-Trees-for-Supervised-Machine-Learning.

[6] Imagen ciclo cvfdt. https://studylib.net/doc/9462259/cvfdt-algorithm.

[7] A. Ben-David. Learning and classification of monotonic ordinal concepts. *Computational Intelligence*, 5:45–49, 1989.

[8] A. Ben-David. Monotonicity maintenance in information-theoretic machinie learning algorithms. *Machine Learning. Issue 1*, Vol.19:29–43, 1995.

[9] J.R. Cano, P.A. Gutiérrez, B. Krawczyk, M. Wozniak, and S. García. Monotonic classification: An overview on algorithms, performance measures and data sets. *Neurocomputing*, Vol.341:168–182, 2019.

[10] J. Gama. A survey on learning from data streams: current and future trends. *Progress in Artificial Intelligence*, Vol. 1:45–55, 2012.

[11] S.A. Jadhav and S.P. Kosbatwar. *Concept-adapting very Fast Decision Tree with Misclassification Error*, volume 5. IJARCET, 2016.

[12] W. Kotlowski and R. Slowinski. On nonparametric ordinal classification with monotonicity constraints. *Progress in Artificial Intelligence*, Vol. 25:2576–2589, 2013.

[13] R. Lior and O.Z. Maimon. *Data Mining With Decision Trees: Theory And Applications*, volume 69 of *Series In Machine Perception And Artificial Intelligence*. World Scientific Publishing Co. Pte. Ldt, 2007.

[14] C. Manapragada, G. I. Webb, and M. Salehi. Extremely fast decision tree. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1953–1962, New York, NY, USA, 2018. ACM.

[15] B. Pfahringer, G. Holmes, and R. Kirkby. New options for hoeffding trees. *Lecture Notes in Computer Science. AI 2007:Advances in Artificial Intelligence*, Vol. 4830:90–99, 2007.

[16] J. Sá, A. Almeida, B. Pereira da Rocha, M. Mota, J.S. De Souza, and L. Dentel. Lightning forecast using data mining techniques on hourly evolution of the convective available potential energy. *Brazilian Congress on Computational Intelligence, Fortaleza*, pages 1–5, 03 2016.