Revista Internacional Tecnologías en la Educación, 5(2), 2018, ISSN 2386-8384

ENSEÑANDO BIG DATA CON LÁPIZ, PAPEL Y TIJERAS

Teaching Big Data With Pen, Paper and Scissors

JUAN FERNANDO SÁNCHEZ-RADA, OSCAR ARAQUE, ÁLVARO CARRERA BARROSO,
CARLOS ÁNGEL IGLESIAS FERNÁNDEZ

Universidad Politécnica de Madrid, España

KEY WORDS

Higher Education
Technologies in Education
Big Data
Artificial Intelligence
Machine Learning

ABSTRACT

This work proposesan approach that combines teaching general concepts in a technology-agnostic fashion with a cooperative learning approach oriented to a the resolution of a challenge in a competitive environment. In this way, students both learn the theory and then put in practice these concepts in class, exploring different options and cooperating in smalls groups. Such groups compete between them through in order to obtain the better solution. Our experience applying this approach in the classroom have been successful. Student satisfaction, test performance, and student understanding are high.

PALABRAS CLAVE

Educación superior
Tecnologías en la educación
Minería de datos, inteligencia artificial
Aprendizaje automático

RESUMEN

Este trabajo propone un enfoque al aprendizaje de Big Data, que combina los conceptos generales de una manera agnóstica a la tecnología, y la puesta en práctica de estos conceptos mediante aprendizaje cooperativo orientado a la resolución de un reto en un entorno competitivo. De esta manera, los alumnos aprenden los conceptos teóricos y los ponen en práctica explorando diferentes opciones y cooperando en grupos. Estos grupos compiten entre sí para obtener la mejor solución. Nuestra experiencia aplicando este enfoque ha sido un éxito.La satisfacción de los estudiantes, el rendimiento y la comprensión de los conceptos son altos.

GLOBAL KNOWLEDGE ACADEMICS

Introducción

El término *Big Data*, también conocido como macrodatos, datos masivos o inteligencia de datos es un concepto amplio, y que hace referencia a un paradigma de captura, tratamiento, transmisión, almacenamiento y visualización de un gran número de datos. Esta definición se esculpe con la ayuda de lo que se conocen como las *tres Vs* (Hashem, 2015): volumen, velocidad y variedad. En lo referente al volumen, un estudio da luz mostrando que, en el 2012, el conjunto total de datos del mundo digital asciende a los 2,7 zettabytes (10^{21}), pudiendo duplicar esta cifra anualmente (Sagiroglu, 2013). Por otro lado, se habla de velocidad en toda la cadena de procesado de los datos, ya que debido a esta ingente cantidad de información, hay procesos que poseen requisitos temporales estrictos, obligando a aumentar la velocidad de los mismos (Zikopoulos, 2011). Por último, la característica de variedad hace referencia a la gran variedad de tipos de datos que se emplean, y que pueden provenir de una miríada de dispositivos o procesos distintos (O'Leary, 2013). En estos tipos cabe destacar los *smartphones*, sensores, redes sociales, fuentes de audio, imagen y vídeo, etc.

Considerando estos matices iniciales, podemos construir una clasificación de los distintos procesos que se desarrollan en el contexto del *Big Data*. Una clasificación extendida en el campo se encuentra en el trabajo de Hashem (2015), que centra su taxonomía en cinco categorías generales que, a su vez, contienen subcategorías más precisas:

- Fuentes de datos, considerando en esta categoría distintas fuentes de datos (proveniente de la web, de redes sociales, de máquinas, de sensores, de actividades humanas, y de *Internet of Things* (*IoT*)). Aquí podemos considerar, además, las estrategias de captura de estos tipos de información.
- Formato de datos. Esta clasificación es general, refiriéndose a tres subtipos utilizados de manera recurrente: estructurados, semi-estructurados y no estructurados.
- Almacenamiento de datos, es decir, las estrategias de almacenamiento de datos. En este campo se distinguen cuatro técnicas, según la estrategia de almacenamiento se base en documentos, columnas grafos o pares clave-valor.
- Normalización de datos. Este paso suele requerir un esfuerzo importante por parte de actores humanos, y su objetivo principal es limpiar, normalizar (estandarizar a un formato común) y transformar los datos de entrada.
- Procesado de datos. Este paso es el más relevante para el trabajo presentado. En esta clase encontramos dos tipos principales de procesado: en tiempo real, o agrupado.

En el presente trabajo nos centramos en el procesado de datos a gran escala, y más concretamente en el procesado agrupado de datos.

Aprendizaje Analógico

La primera fase del aprendizaje consiste en la adquisición de los conceptos relacionados con las técnicas de *Big Data* y aprendizaje automático. Para ello, se dividen los conocimientos en dos temas: despliegue de arquitecturas de Big Data, y aplicación de aprendizaje automático en *Big Data.*

Para el primer tema, se combinan dos métodos de enseñanza. Primero, se explican los conceptos básicos desde un punto de vista teórico utilizando transparencias en clase y recursos adicional de lectura opcional. Al terminar el tema, se han cubierto todos los conceptos básicos, desde el particionado de datos para distribuir el procesamiento, hasta el funcionamiento de orquestadores de procesamiento distribuido. En particular, se explican los algoritmos básicos para gestionar el procesamiento, y los actores que toman parte en el procesamiento. Aunque estos conceptos son importantes y comunes (con ciertas peculiaridades) a cualquier arquitectura de procesamiento de *Big Data*., las plataformas de análisis suelen abstraerlos estos detalles de bajo nivel. Sin embargo, estos detalles tienen un impacto muy grande en el rendimiento del análisis. Conocer el funcionamiento interno de la plataforma puede llevar a mejoras de órdenes de magnitud en la velocidad de procesamiento.

Para interiorizar estos conceptos, al finalizar la parte teórica, se realiza un ejercicio de simulación de procesamiento de procesamiento de datos "analógico". El objetivo es realizar una tarea de procesamiento distribuido (p.e. contar el número de ocurrencias de cada palabra en un texto). Para ello, se cuenta con los datos a analizar en formato físico (p.e. una hoja de papel con el texto a analizar). Primero, la clase en conjunto llega a un acuerdo con ayuda del profesor sobre las operaciones a alto nivel que han de realizarse sobre los datos. Seguidamente, cada alumno recibe un rol, que coincide con uno de los elementos software presentes en la plataforma de *Big Data*. Por ejemplo, estas tareas pueden ser: particionar los datos de entrada (cortar la hoja de papel en trozos más pequeños que contengan sólo una frase), distribuir la información entre los "procesadores" disponibles (repartir papeles a grupos de alumnos), procesar cada trozo de información (contar palabras en cada frase), agregar los cálculos independientes (sumar cada cuenta), etc.

El segunda tema consiste en la aplicación de aprendizaje automático utilizando herramientas software conocidas. En este caso, se aplica una combinación de clases teóricas presenciales con prácticas guiadas mediante *Jupyter Notebooks* (Jupyter, 2018), una tecnología que permite distribuir código interactivo con explicaciones con texto e imágenes. De esta forma, los alumnos aprenden los conceptos básicos y los aplican de forma incremental en un entorno controlado, sobre un conjunto de datos conocido. Estos conocimientos prácticos les serán necesarios para la parte de aprendizaje cooperativo, que además les permitirá poner a prueba sus habilidades y ampliarlas mediante un reto.

Aprendizaje Cooperativo

En este trabajo proponemos el uso de técnicas de ***aprendizaje cooperativo*** en la enseñanza técnica de *Big Data* y **Aprendizaje Automático**. El aprendizaje cooperativo es un tipo de aprendizaje colaborativo en el que los alumnos trabajan en equipos con tareas de aprendizaje estructurado bajo una serie de condiciones (Millis, 1997):

- Interdependencia positiva, ya que los miembros del mismo equipo deben depender unos de otros para llegar a un objetivo.
- Responsabilidad individual. Los miembros de un equipo se hacen responsables de hacer su parte de trabajo y, además, de conocer todo el material relacionado con el mismo.
- Uso de habilidades interpersonales. Los miembros del equipo practican habilidades relacionadas con liderazgo, toma de decisiones, comunicación, y manejo de conflictos.
- Interacción cara a cara. El trabajo se realiza, por lo general, con los miembros del equipo trabajando físicamente juntos.
- Evaluación del funcionamiento del grupo. Los equipos, con cierta frecuencia, reflexionan acerca del trabajo que están desarrollando, al igual que cómo podrían mejorarlo o que acciones pueden tomar en el futuro.

Además, para completar la definición y las características del aprendizaje cooperativo, caben destacar tres niveles del mismo (Johnson, 1998): aprendizaje cooperativo informal, aprendizaje cooperativo formal, y grupos cooperativos. Éstos últimos son una forma menos común de este tipo de aprendizaje en el que los grupos permanecen juntos con el ánimo de ofrecerse apoyo mutuo tanto académico como personal. Este nivel de aprendizaje cooperativo se da en la academia.

Respecto a los otros dos niveles, la diferencia entre aprendizaje cooperativo informal y formal es más sutil. Mientras que en el informal los estudiantes se juntan en grupos para realizar tareas concretas durante no más del periodo de una clase; en la variante formal esta agrupación puede llegar a durar varias semanas, por lo que se puede aplicar a la duración de una asignatura universitaria (Johnson, 1998).

Implantar el aprendizaje cooperativo en el aula conlleva varias ventajas (Millis, 1997). Entre otras, la interacción entre alumno-alumno y alumno-profesor se facilita y agiliza; las calificaciones suben, al igual que la capacidad de retención de los alumnos, la motivación hacia el temario y la actitud hacia la asignatura. Además, las capacidad de razonamiento de los alumnos se mejoran, así como las capacidades de comunicación, de trabajo en equipo y de relaciones interpersonal. Por último, se aumento la autoestima y se reduce el nivel de ansiedad entre el alumnado.

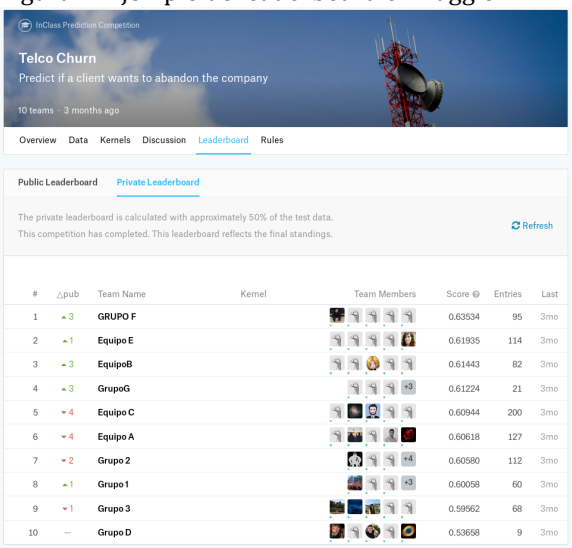
Kaggle: herramienta para el aprendizaje cooperativo

Kaggle es una plataforma utilizada para competiciones de modelado predictivo y analítico en la que estadísticos, científicos y mineros de datos compiten para producir los mejores modelos de predicción y descripción de los conjuntos de datos cargados por empresas y usuarios. Este enfoque de *crowdsourcing* se basa en el hecho de que existen innumerables estrategias que se pueden aplicar a cualquier tarea de modelización predictiva y es imposible saber de antemano qué técnica o análisis será más eficaz.

Actualmente, hay más de 83.000 usuarios en activo en la plataforma1(Kaggle, 2018), siendo ésta la herramienta más popular en el campo de la minería de datos. Es importante señalar que Kaggle permite el procesado de datos a gran escala, pudiendo incluir dicha plataforma en el ecosistema del *Big Data*. Las capacidades que Kaggle ofrece, junto a su facilidad de uso, hace de esta plataforma una herramienta ideal para su uso en la clase. De hecho, esta aplicación se ha hecho tan popular en universidades de todo el mundo que se ofrece el servicio gratuito *Kaggle In Class*, específicamente destinado a los profesores pueden preparar y proponer competiciones a sus alumnos.

En el marco de una competición, Kaggle ofrece tanto a profesores como alumnos todas las herramientas para el correcto desarrollo de la actividad. Así, los profesores pueden preparar y alojar los datos de la competición, publicándolos para su uso durante la misma. Además, la evaluación de los trabajos de los alumnos se realiza de forma automática, en función de la calidad de la resolución del problema propuesto. Para los alumnos, Kaggle contiene todos los datos necesarios para enfrentarse al problema, así como un sistema centralizado al que enviar sus soluciones.

Figura 1: Ejemplo de leaderboard en Kaggle



Una parte especialmente relevante de estas competiciones *In Class* es la posibilidad de organizar los resultados de cada grupo de alumnos en un *leaderboard* público, de manera que la calidad de las soluciones se muestra tanto a profesores como alumnos. Un ejemplo de esta característica se puede apreciar en la Figura 1. Como se puede observar, a cada equipo de alumnos se le corresponde una puntuación (*score*) que indica la calidad de la solución aportada. Resulta especialmente relevante el hecho de que estas puntuaciones sea públicas y consultables, ya que los alumnos conocen en todo momento la calidad de sus soluciones, así como las de otros equipos. Creemos que es importante sopesar la importancia de este hecho, y de cómo puede afectar a los alumnos, y al desarrollo de la actividad. Esta cuestión se aborda más detalladamente en la siguiente secciones.

Evaluación

En el año 2018 se ha impartido, por tercer año consecutivo, la asignatura de *Sistemas de información y tecnologías del conocimiento*, en la Universidad Politécnica de Madrid, en la titulación de Máster Universitario en Ingeniería de Telecomunicación. Esta asignatura, enmarcada en el segundo año de un máster habilitante comprende, entre otros, contenidos de procesado de datos y aprendizaje automático.

Durante dicho curso se ha puesto en práctica la combinación de aprendizaje analógico con aprendizaje cooperativo mediante la propuesta de una competición en Kaggle que se denomina *Telco Churn* (Telco-churn, 2018). En esta actividad, los alumnos, organizados en grupos de cinco miembros, deben poner en práctica sus habilidades de procesado de datos y aprendizaje automático para predecir si un cliente de una empresa de telecomunicaciones va a abandonar dicha compañía o no. La actividad se plantea para su desarrollo durante tres semanas en las que los alumnos diseñan, implementan y evalúan sus métodos predictivos. Como se ya se ha comentado, la plataforma Kaggle facilita el desarrollo de esta actividad tanto a profesores como a alumnos.

Con ánimo de evaluar la eficacia y la acogida de la actividad basada en una competición, se han realizado entrevistas con todos los alumnos participantes. Estas actividades consistieron en preguntas acerca de la actividad, seguidas de comentarios aportados por los alumnos. Adicionalmente, se plantean tres preguntas en las que se pide a los alumnos que muestren su acuerdo con tres afirmaciones, que mostramos a continuación.

1. La participación en la competición me ha resultado estimulante.
2. Considero que con mi participación en la competición he mejorado mis capacidades.
3. En general, estoy satisfecho con la competición.

De manera general, la retroalimentación obtenida de los alumnos resulta positiva. El detalle de las respuestas, agregado como el porcentaje sobre el número total de alumnos, se muestra en la siguiente tabla.

	En desacuerdo	Neutral	De acuerdo
Afirmación 1	18 %	25 %	57 %
Afirmación 2	14 %	18 %	68 %
Afirmación 3	16 %	25 %	59 %
Media valores	16 %	23 %	61 %

Como se puede ver, el porcentaje de alumnos que está en desacuerdo con las afirmaciones propuestas no supera, en ningún caso, el 18%, y su media se encuentra en el 16%. Consideramos dicho resultado enormemente positivo, indicando que los alumnos han experimentado al menos algunas de las ventajas que el aprendizaje cooperativo conlleva. Respecto a la fracción de alumnos que se muestra de acuerdo con las afirmaciones, la media se eleva al 61%, con valores máximos de 68% y mínimos del 57%. De nuevo, estos resultados resultan alentadores, significando que, aunque el método implementado tiene un margen de mejora, en general se puede considerar satisfactorio en su ámbito.

Además, las percepción general de los estudiantes, expresadas a través de las entrevistas, es que la propuesta de la actividad cooperativa resulta atractiva para los alumnos, y les motiva para explorar el contenido impartido en la asignatura. Como último comentario, los alumnos también han manifestado una actitud positiva respecto al carácter competitivo de la actividad, argumentando que el hecho de tener que competir con otros equipos de alumnos ha aumentado de manera

relevante la motivación hacia la asignatura, e incluso ha animado a los alumnos a buscar material adicional que no es impartido en asignatura.

Conclusiones

El aprendizaje cooperativo es una valiosa herramienta que, aunque ampliamente estudiada, no se implanta con suficiente frecuencia en los planes de estudio de titulaciones superiores. En este trabajo se destaca el importante papel que este tipo de aprendizaje puede tener en la enseñanza de temáticas técnicas relacionadas con la minería de datos y el aprendizaje automático. En su versión más completa, este trabajo propone la implantación de una estrategia de aprendizaje colaborativo que lleva tanto a docentes como estudiantes a embarcarse en un proceso de investigación que pasa desde la propuesta de un problema que permite a los alumnos poner en práctica los contenidos del curso, hasta la resolución colaborativa en equipos, pasando por un proceso iterativo de exploración de opciones de resolución en el que los docentes pueden y deben implicarse.

Por otro lado, la experiencia con la simulación de procesamiento analógico ha sido limitada, pero muy positiva. Los comentarios de los alumnos indican El párrafo de arranque no tiene sangría.

Los párrafos segundo y siguientes deben tener una sangría de primera línea de 0,5.

que les permite comprender los conceptos de forma aplicada, a nivel conceptual, pero sin la abstracción de los típicos sistemas de procesamiento *Big Data.*

Se han presentado resultados empíricos que indican que la implantación de la metodología propuesta cuenta con una amplia acogida por el alumnado. Además, a través de entrevistas con estudiantes se ha visto que una combinación de aprendizaje cooperativo y competitivo resulta altamente atractiva para el alumnado, aumentado el nivel de entrega de éstos hacia la asignatura. Es preciso observar que, en comparación a años anteriores en los que el entorno cooperativo y competitivo no se ha implementado, las calificaciones, en general, son más altas.

Como trabajo futuro, creemos que un estudio más extenso acerca de la implantación de un paradigma cooperativo y competitivo sería interesante, con especial énfasis en el equilibrio entre ambos extremos. Adicionalmente, es importante destacar que este tipo de estrategias pueden ser adaptadas a un gran número de campos de conocimiento técnico, como puede ser la ciberseguridad, el diseño de sistemas, o la planificación de procesos, entre otros.

Referencias

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98-115.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems*, *28*(2), 96-99.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). *Active learning: Cooperation in the college classroom*. Interaction Book Company, 7208 Cornelia Drive, Edina, MN 55435.

Millis, B. J., & Cottell Jr, P. G. (1997). *Cooperative Learning for Higher Education Faculty. Series on Higher Education*. Oryx Press, PO Box 33889, Phoenix, AZ 85067-3889.

Jupyter, http://jupyter.org/. Último acceso el 10 de Julio de 2018.

Kaggle. https://www.kaggle.com/. Último acceso el 10 de Julio de 2018.

Telco-churn. https://www.kaggle.com/c/telco-churn/. Último acceso el 10 de Julio de 2018.