Análisis del abandono estudiantil basado en la interacción*

Realizado a través de un proceso de *Knowledge Discovery on Databases*

Alfonso Barragán Carmona
Escuela Superior de Informática
Ciudad Real, Castilla la Mancha
alfonso.barragan@alu.uclm.es

Javier Monescillo Buitron
Escuela Superior de Informática
Ciudad Real, Castilla la Mancha
javier.monescillo@alu.uclm.es

Roberto Plaza Romero
Escuela Superior de Informática
Ciudad Real, Castilla la Mancha
roberto.plaza@alu.uclm.es

ABSTRACT

Actualmente la mayoría de usuarios hacen uso de la tecnología para adquirir nuevos conocimientos, una de las formas más sencillas de poder hacerlo es la compra de un curso *online*, pero ¿Realmente existe una eficiencia a la hora de aprender en un curso de este tipo? ¿Este tipo de cursos son realmente buenos?

Los cursos *online* se han convertido en toda una tendencia para los usuarios comunes de Internet, pero también se ha convertido en una tendencia el abandono de los mismos. En el presente trabajo se ha elaborado un sistema basado en minería de datos para evitar el abandono de cursos *online*, para la realización de dicho sistema, se trazará un proceso de *Knowledge Discovery on Databases*, abreviando, un proceso KDD. El sistema tratará de mejorar la tasa de abandono en los cursos, este tiene una limitación, y es que solo se estudiará bajo la perspectiva del curso, no la del usuario.

Para este estudio en particular se ha tomado como metodología principal estudiar por entidades dentro del escenario, centrándonos en la idea/concepto de curso, pero no olvidando los módulos y los usuarios.

El proceso KDD constará de los siguientes pasos: **Creación de los datos objetivo**, **Función de la minería de datos**, **Elección del algoritmo a aplicar**, **Búsqueda de patrones**, **Transformación**, **Resultados**, **Interpretación** y **Conclusión**.

Como extracto principal de las conclusiones, se puede decir que la mayoría de estudiantes de dichos cursos abandonaron con una tasa del 90%, por lo cual, el sistema es necesario para poder mejorar la tasa y los cursos.

KEYWORDS

Data mining, KDD, educación, online, cursos, formación

ACM Reference Format:
Alfonso Barragán Carmona, Javier Monescillo Buitron, and Roberto Plaza Romero. 2018. Análisis del abandono estudiantil basado en la interacción: Realizado a través de un proceso de *Knowledge Discovery on Databases*. In *Proceedings of ACM Congreso olivas (SMILESTOCK'18).* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.475/123_4

*Produces the permission block, and copyright information

SMILESTOCK'18, Diciembre 2018, Ciudad Real, Castilla la Mancha ESPAÑA

ACM ISBN 123-4556-789/08/06.
https://doi.org/10.475/123_4

1 INTRODUCCIÓN

El problema a tratar es la segmentación de módulos de un determinado curso online, que provoca que los usuarios abandonen o dejen de lado el aprendizaje. Los datos que estan disponibles, son extraídos de *MOOC.es*[1], al tratarse de un recopilador de cursos, no se cuenta con unos datos detallados sobre el perfil de los usuarios o los cursos en si mismos, pero contiene unos registros detallados sobre la interacción de los usuarios con cada uno de los elementos de un curso, debido a esto, se plantea la hipótesis de que existe un patrón en la actividad que tienen los usuarios con los diversos módulos, que al final desencadena el abandono o no.

2 PROCESO DE RESOLUCIÓN DEL PROBLEMA

El método tradicional de manejo de datos se ha quedado obsoleto, y es impráctico, sobre todo en grandes volúmenes de datos, ya que en bases de datos y otros contenedores de datos, el volumen de archivos y su capacidad está creciendo exponencialmente.
Por ello, el verdadero valor de estos datos recae en la capacidad del usuario para extraer de ellos informes, análisis estadístico e inferencia.
El término KDD se utiliza para obtener conocimiento útil de los datos, la minería de datos es un paso particular en este proceso, es la aplicación de unos algoritmos específicos para la extracción de patrones.
El proceso KDD es interactivo e iterativo, este proceso tiene varios pasos, para el caso de estudio trazamos la siguiente hoja de ruta.

2.1 Dominio de la aplicación

2.1.1 Alcance. El dominio de la aplicación es la mejora de los cursos vía telemática, es decir, mejorar la precisión de dichos cursos, para disminuir la cantidad de usuarios que los abandonan, mirando siempre desde la perspectiva del curso

2.1.2 Límites. La limitación del sistema está en el tratamiento de cursos, ya que no clasificamos a los alumnos, es decir, no hacemos ningún tipo de perfil en función del abandono.

2.2 Creación de los datos objetivo

Para emular la experiencia de un proceso KDD moderno, la primera tarea, en concreto, fue volcar el contenido de los ficheros *csv* provistos, a una base de datos MySQL.

[1]Pagina web que registrascursos de diversas paginas como *edx.org* o *coursera.org*

Tras esta migración de formato, comenzamos un proceso de pruebas en paralelo a un proceso de deliberación sobre los datos que fueran de relevancia, al ser todos los datos de tipo categórico o de tipo tiempo, se decide comenzar el siguiente proceso de reducción de datos.

2.2.1 Reducción en los datos. En cuanto a la reducción de los datos, en el problema solo se ha estudiado desde la perspectiva de los cursos, partiendo del uso de la primitiva **SUM** de *MySQL*, sintetizando así los registros de resumen y consiguiendo transformar las variables categóricas. Además se realizará un proceso de normalizado, por el cual, se procederá a la obtención del % de abandono.

2.2.2 Limpieza de datos. Particularmente el problema de la KDD Cup de 2015 solo tiene variables categóricas, por lo que realmente la limpieza en los datos no es necesaria, no existen valores nulos, ni registros vacíos entonces no se requiere de realizar un filtrado de registros excesivo.

2.2.3 Contenido de los datos objetivo.

- Id curso
- Número total de usuarios
- Número de submódulos por tipo por curso
- Número total de submódulos
- Número de interacciones por tipo por curso
- Número total de interacciones por curso
- Número de abandonos
- Número de no abandonados
- Porcentaje de éxito

2.3 Función de la minería de datos

La función principal de la *minería* de datos en el problema a tratar, es focalizar el esfuerzo en encontrar los tres principales tipos de cursos y clasificarlos, para ello, se empleará una búsqueda de patrones en función de la tasa de abandonos/curso.

Se cuenta con los siguientes tipos de curso:

- Curso de calidad
- Curso mejorable
- Curso de mala calidad

2.4 Elección del algoritmo a aplicar

Tras diversas deliberaciones entre los miembros del equipo se ha optado por la elección de un árbol de decisión, ya que estamos buscando características dentro de los datos objetivo.

Buscamos, no sólo encontrar patrones si no, también clasificar nuevos sucesos, para ello los árboles de decisión y modelos *Random Forest* son especialmente útiles, razón por la cual nos decantamos por el árbol

Un árbol de decisión además de útil para esta tarea cuenta con otras bondades, por ejemplo la posibilidad de paralelizar la ejecución y que esta no tome un tiempo excesivo.

Además susodicho modelo solventa la, a veces, complicada tarea de representar, no solo información, sino también conocimiento.

2.5 Búsqueda de patrones

A la hora de realizar la búsqueda de patrones, se tiene que realizar una ordenación de los cursos en función de la tasa de abandono, en el siguiente caso, una ordenación de mayor a menor respecto a la tasa de abandono.

Se ha optado por la división de todos los cursos en tres grupos igual-

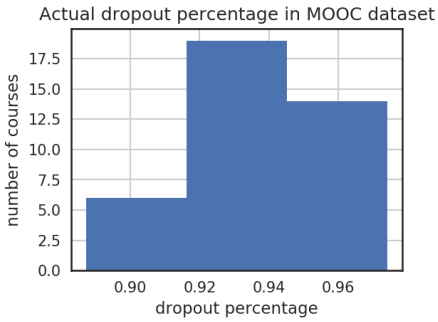

Figure 1: Grupos en función del tipo de curso

mente representados, por ende, la definición de curso de alta calidad, curso fronterizo y baja calidad, puede verse muy afectada. Estas pautas han podido afectar a la hora de descubrir dichos patrones.

2.6 Transformación

Para ayudar a los encargados de aquellos cursos con mayor tasa de abandono se ha provisto al sistema de una característica que transforma el árbol de decisión generado a código Python haciendo uso de librerías que encontramos en stackoverflow.

Lamentablemente en esta primera versión del sistema el operador ha el encargado de notificar los resultados de el aprendizaje y aplicar alguna corrección si así se requiere. Por fortuna el lenguaje de salida es código Python, comprensible para cualquier persona que sepa inglés.

3 RESULTADOS

Los resultados obtenidos tras la realización del proceso KDD, son aceptables, tras todo el preproceso y construcción del modelo se ha obtenido en siguiente árbol de decisión.[Figure 2]

Tras obtener una precisión del 72%, se comprueba en *big ML* la precisión obtenida por el mejor de sus modelos, llegando a un 85%.[Figure 3]

4 INTERPRETACIÓN

4.1 ¿Cómo interpretamos los datos?

Los datos, o por lo menos como a nivel personal los entendemos dejan un atisbo agridulce en nuestras bocas. Siempre se ha dicho que no hay mal alumno sino horrible maestro, más estos datos tan sólo dejan ver lo contrario.

Como ya hemos aclarado antes, la idea de realizar un árbol de decisión es bastante acertada puesto que el resultado es muy visual.

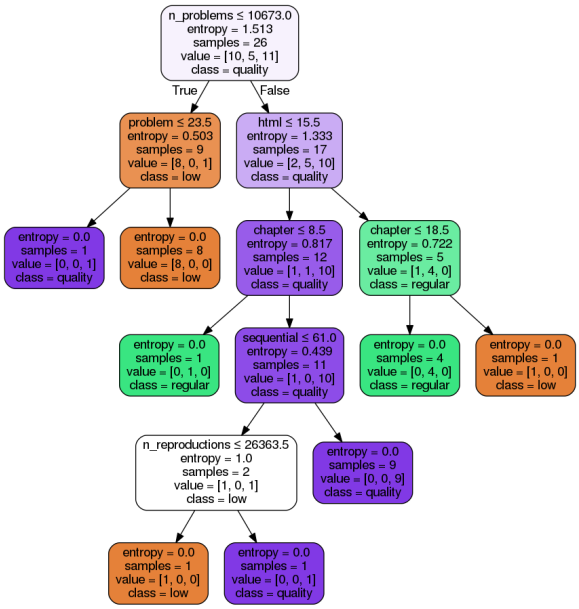

Figure 2: Árbol de decisión

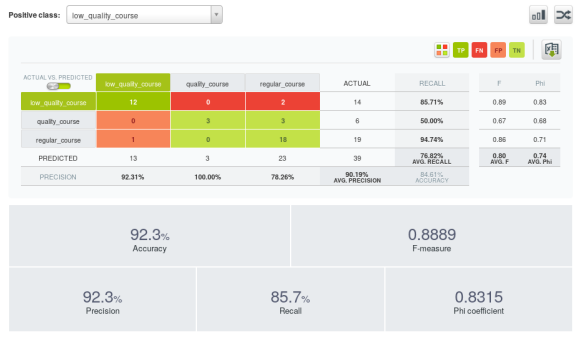

Figure 3: Reporte en big ML

Podemos ver que de cuantos más ejercicios consta un módulo más posibilidades hay de que el curso al que pertenece no caiga en el olvido.

Los límites, como podemos ver están muy poco claros puesto que apenas hay una diferencia mayor al 10% entre el mejor y el peor de los cursos, dando lugar a escenarios bastante confusos.

5 CONCLUSIONES

Como primera conclusión ha de remitirse al último apartado, dejando claro que el factor más importante, con cerca del 60% de peso a nivel teórico, es el número de ejercicios. Nos gustaría pensar que por que estos mantienen a los pupilos ocupados el tiempo suficiente para acabar el curso.

Análogamente el primer vistazo al porcentaje de abandonos deja claro que estos cursos no son algo serio, por lo menos a ojos de los alumnos. Por ende un índice de abandono de o en torno al 90% no es un caso aislado, sino un caso, inclusive bueno.

Como conclusión final dejar constancia del error de selección de entidad que estudiar. Centrarse en los cursos parecía muy prometedor, más el problema parecía estar en otro lugar. Completar un curso parece tener más que ver con la intención del alumno que con el contenido del curso.

6 TRABAJO FUTURO

Una vez concluido el trabajo, queda preguntarse: ¿Cuán puede esto dar de sí?

En primer lugar, nos gustaría que la fase de transformación fuese menos rudimentaria, es decir, que la manera de avisar al supuesto operador del sistema vaya más allá que una línea de texto. La salida tan simple parece un desperdicio.

Después, aunque no menos importante, estudiar la entidad de los alumnos puede ser muy interesante. Desafortunadamente el estudio de cursos se hace cuesta arriba, ya que la mayoría son muy similares, por ello la idea de buscar un prototipo de alumno que termina el curso puede ir más encaminada.

Por último, el problema nos ha gustado por lo cual se podría ampliar el dominio de tal manera que pudiéramos tratar datos de plataformas menos propensas al abandono, como puedan ser moodle y udemy.

REFERENCES

Material misceláneo ofrecido durante la asignatura *Minería de Datos* de la Escuela Superior de Informática de Ciudad Real.

A destacar el artículo *The KDD Process for Extracting Useful Knowledge from Volumes of Data* escrito por U.Fayyad, G.Piatetsky-Shapiro y P.Smyth.

Diferentes entradas en www.stackoverflow.com