XXI Encuentro Internacional de Educación a Distancia
Educación virtual en los cinco continentes
Del 2 al 6 de diciembre de 2013
ISSN: 2395-8901

# IDENTIFICACIÓN Y COMPRENSIÓN DE LOS USOS DE LAS TIC

**Identificación y comprensión de los usos de las TIC por parte de estudiantes universitarios a distancia en sus procesos de aprendizaje empleando minería de datos**

Dr. Alejandro Canales Cruz1, Dr. Jesús Humberto González González2

1 Coordinación de Universidad Abierta y Educación a Distancia de la Universidad Nacional Autónoma de México

alejandro_canales@cuaed.unam.mx

2 Facultad de Psicología de la Universidad Autónoma de Nuevo León

jesushumbertogonzalez@gmail.com

**Reseña**

En este trabajo se propone la aplicación de técnicas de minería de datos para identificar los usos de TIC por parte de estudiantes universitarios de la modalidad a distancia. Se utilizarán los datos reales

en 393 estudiantes de diversas instituciones educativas de México y se emplearán métodos de clasificación de caja blanca, como las normas de inducción y árboles de decisión. Los experimentos intentan mejorar su precisión para identificar patrones de usos de TIC de los estudiantes para mejorar el diseño de las situaciones de aprendizaje en contextos formales a distancia, haciéndolas más próximas, adaptadas y motivadoras para éstos, teniendo en cuenta los diferentes perfiles de estudiantes identificados.

**Introducción**

Nuestro mundo se encuentra abrumado con datos. La cantidad de datos en el mundo y en nuestras vidas parece cada vez más y continúa incrementándose. Por una parte, los equipos de cómputo cuentan con una capacidad de procesamiento cada vez mayor, un sistema almacenamiento que permite guardar prácticamente cualquier cantidad de información que manejemos a nivel personal, son equipos portables y tienen acceso a Internet. Por otra, la World Wide Web (WWW), nos abruma con la gran cantidad de información que nos ofrece. Todos podríamos dar testimonio de la creciente brecha entre la generación de los datos y el conocimiento de la misma. Sin embargo, a medida que el volumen de datos aumenta, inexorablemente, la proporción de lo que la gente entiende disminuye alarmantemente.

En el campo de la minería de datos, los datos se almacenan electrónicamente y se automatiza su la búsqueda, a través del uso de equipo de cómputo. Es un área que no es nueva y que ha sido aplicada en áreas del conocimiento, tales como la educación, economía, estadística, meteorología, ingeniería, etc., con la intención de buscar patrones en los datos que permitan de forma automática identificar y validar predicciones de un fenómeno dado. Lo nuevo de esta disciplina radica en el espectacular aumento de las oportunidades para encontrar patrones en los datos. Mientras el mundo crece en

complejidad y nos abruma con los datos que se generan, la minería de datos se convierte en una gran opción para dilucidar patrones ocultos. El análisis inteligente de los datos es un recurso valioso que puede conducir a nuevos conocimientos y ventajas. De acuerdo con Aluja (2001), la utilización de la minería de datos con respecto a los modelos estadísticos se puede diferenciar:

- La minería de datos es un proceso completo formado por varias etapas, tales pre-procesado, la aplicación de técnicas de minería de datos (una de ellas puede ser estadística) y la evaluación e interpretación de los resultados.
- En minería de datos se lleva a cabo un análisis utilizando el porcentaje de datos bien clasificados. Mientras que en las técnicas estadísticas el análisis de datos se realiza con base en la verosimilitud de los datos dado el modelo.
- En minería de datos se suele utilizar una búsqueda basada en meta-heurísticas mientras que en estadística la búsqueda suele realizarse mediante la modelización basada en un algoritmo de ascenso de colinas (conocido en inglés como hill-climbing) en combinación con un test de hipótesis basado en razón de verosimilitud.
- La minería de datos está orientada a trabajar con cantidades muy grandes de datos (millones y billones de datos). En cambio la estadística no suele funcionar tan bien en bases de datos de tan gran tamaño y alta dimensionalidad.

Una solución a identificar y comprender los usos de las TIC por parte de estudiantes universitarios a distancia en sus procesos de aprendizaje, es el uso de técnicas de extracción de conocimiento o minería de datos en educación, lo que ha dado lugar a la denominada minería de datos educativa (Educational Data Mining, EDM por su siglas en inglés) (Romero

> y Ventura, 2007). Esta nueva área de investigación se ocupa del desarrollo de métodos para explorar los datos que se dan en el ámbito educativo, así como de la utilización de estos métodos para entender mejor a los estudiantes y los contextos en que ellos aprenden (International Educational Data Mining Society, 2013).

Las técnicas de EDM ya se han empleado con éxito para crear modelos de predicción del rendimiento de los estudiantes (Kotsiantis, Patriarcheas y Xenos, 2010), obteniendo resultados prometedores que demuestran cómo determinadas características sociológicas, económicas y educativas de los alumnos pueden afectar en el rendimiento académico.

Es importante también destacar que hasta la fecha la mayor parte de las investigaciones sobre minería de datos se aplicada a los problemas identificados en programas educativos pertenecientes a la modalidad presencial, se han aplicado, sobre todo, en el nivel de educación superior (Kotsiantis, 2009) y, en menor medida en la modalidad de educación a distancia (Lykourentzou, Giannoukos, Nikolopoulos, Mpardis y Loumos, 2009).

En este trabajo se propone la utilización de técnicas de minería de datos para detectar los usos de las TIC por parte de estudiantes universitarios a distancia en sus procesos de aprendizaje. Para ello, se propone utilizar diferentes técnicas de minería de datos debido a que es un problema complejo, los datos suelen presentar una alta dimensionalidad (hay muchos factores que pueden influir). El objetivo final es identificar elementos que nos permitan mejorar el diseño de las situaciones de aprendizaje en contextos formales a distancia, haciéndolas más próximas, adaptadas y motivadoras para éstos, teniendo en cuenta los diferentes perfiles de estudiantes identificados y aprovechando la

identificación y uso que estos muestran por determinados recursos tecnológicos. En suma, en futuro cercano se tendrán las bases para construir un modelo de aprendizaje que servirá para preparar estudiantes con un perfil profesional esencialmente apropiado para la sociedad del conocimiento.

La organización del presente artículo está contemplada de la siguiente forma: En la siguiente sección se describe de manera general el método que se propuso para identificar los usos de las TIC por parte de los estudiantes. Enseguida, se describen los datos que se utilizaron para este trabajo. A continuación se explica las tareas de pre-procesado de los datos que se han llevado a cabo. La siguiente sección describe las diferentes pruebas de minería de datos que se han realizado y los resultados obtenidos de las mismas. Enseguida se realiza una interpretación de los resultados y finalmente se presentan las conclusiones y trabajo a futuro.

Método utilizado para identificar los usos de las TIC

En la figura 1 se muestra el método empleado en la identificación de usos de las TIC de estudiantes universitarios de la modalidad a distancia.

En la etapa de recopilación de datos se obtiene toda la información disponible de los estudiantes. Para ello primero se diseñó un instrumento dividido en nueve dimensiones (Gonzáles y Canales, 2013): datos del participante, pedagógica-didáctica, organizacional, comunitaria, administrativa, empleo didáctico de las TIC, métodos de aprendizaje y enseñanza, aplicación de la tecnología y multimedios TIC.

Para la etapa de Pre-procesado se llevan a cabo acciones como limpieza de datos, transformación de variables, particionado de datos; además, se aplican técnicas como la selección de atributos y el re-

balanceado de datos para intentar solucionar los problemas de la alta dimensionalidad y desbalanceo que presentan normalmente este tipo de conjuntos de datos.

Posteriormente, se aplican las técnicas de minería de datos. En esta etapa se aplican algoritmos para identificar los usos de las TIC por parte de estudiantes universitarios a distancia.

Por último, en la etapa de interpretación de los resultados se analizan los modelos que han obtenido los mejores resultados para utilizarlos en la identificación de usos de las TIC. Para ello, se analizan los factores que aparecen en las reglas de decisión, los valores que presentan y como están relacionados con otros factores.

Recopilación y pre-procesamiento de los datos

Los datos empleados se recopilaron de la aplicación del instrumento antes mencionado (Gonzáles y Canales, 2013). El instrumento se aplicó a estudiantes universitarios de la modalidad a distancia pertenecientes a:

- Maestría en docencia con orientación en educación media superior que pertenecen a la División de Posgrado de la Facultad de Psicología de la Universidad Autónoma de Nuevo León.
- Facultad de Economía de la Universidad Nacional Autónoma de México.
- Doctorado en Sistemas y Ambientes Educativos del Sistema de Universidad Virtual de la Universidad de Guadalajara

A continuación en la tabla 1, se resumen los resultados obtenidos por módulo y las variables que la integran.

En la etapa de pre-procesamiento de los datos se comienza con la limpieza de los mismos. Para ello, se extrajo del conjunto de datos a aquellos estudiantes que seleccionaron la misma respuesta en todas las preguntas. Es decir, que contestaron en todos los casos estar "muy frecuentemente" o "nada frecuente". En la etapa de discretización, se asignó a las respuestas un rangos: Totalmente de acuerdo (10 a 7.5); De acuerdo (7.4 a 5); Ni de acuerdo, ni en desacuerdo (4.9 a 2.5) y En desacuerdo (2.4 a 0).

El software de minería de datos que se ocupó fue Weka, por lo que se creó un archivo con formato .ARFF (Weka, 2013). Después de realizar las anteriores tareas de pre-procesado, se dispone de un primer fichero de datos con 40 atributos sobre 393 estudiantes. Debido a la gran cantidad de atributos recopilados (40), se realizó también un análisis o estudio de selección de atributos para determinar cuáles son los que mayormente influyen en la variable de salida oclase a predecir (usos de las TIC). Para seleccionar las variables de mayor relevancia se utilizaron varios métodos de selección de atributos disponibles en el software Weka. En general, estos algoritmos de selección pueden ser agrupados por varios criterios. Una categorización popular es aquella en la que los algoritmos se distinguen por su forma de evaluar atributos y se clasifican en: filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje y wrappers (envoltorios), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto.

Aplicación de minería de datos y resultados

Se realizaron varios experimentos con el objetivo de obtener la máxima exactitud de clasificación. En un primer experimento se ejecutó 10 algoritmos de clasificación utilizando todos los atributos con los que se cuenta, es decir de toda la información disponible. En un segundo experimento, se utilizó sólo los mejores atributos. Se ha seleccionado 10 algoritmos de clasificación de entre los disponibles por la herramienta de minería de datos Weka. Esta selección se ha realizado debido a que estos algoritmos, son todos del tipo "caja blanca", es decir, se obtiene un modelo de salida comprensible para el usuario, porque o se obtienen reglas de clasificación del tipo "Si – Entonces" o árboles de decisión. De esta forma un usuario no experto en minería de datos como un profesor o instructor puede utilizar directamente la salida obtenida por estos algoritmos para detectar los usos de las TIC de estudiantes universitarios y poder tomar decisiones sobre cómo mejorar los cursos en línea.

Las reglas de clasificación del tipo "Si – Entonces" son una manera simple y fácilmente comprensible de representar el conocimiento. Una regla tiene dos partes, el antecedente y el consecuente. El antecedente de la regla (la parte del "Si") contiene una combinación de condiciones respecto a los atributos de predicción. El consecuente de la regla (la parte del"Entonces") contiene el valor predicho para la clase. De esta manera, una regla asigna una instancia de datos a la clase señalada por el consecuente si los valores de los atributos de predicción satisfacen las condiciones expresadas en el antecedente, y por tanto, un clasificador es representado como un conjunto de reglas. Los algoritmos incluidos en este paradigma pueden ser considerados como una búsqueda

heurística en un espacio de estados. En este caso, un estado corresponde a una regla candidata, y los operadores corresponden a la generalización y especialización de operaciones que transformen una regla candidata en otra. Los algoritmos de inducción de reglas de clasificación que se usaron son: OneR, Ridor, JRip y NNge.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, el cual contiene cero o más nodos internos y uno o más nodos de hoja. Los nodos internos tienen dos o más nodos secundarios y contienen divisiones, los cuales prueban el valor de una expresión de los atributos. Los arcos de un nodo interno a otro secundario (o de menor jerarquía) son etiquetados con distintas salidas de la prueba del nodo interno. Cada nodo hoja tiene una etiqueta de clase asociada. El árbol de decisión es un modelo predictivo en el cual una instancia es clasificada siguiendo el camino de condiciones cumplidas desde la raíz hasta llegar a una hoja, la cual corresponderá a una clase etiquetada. Un árbol de decisión se puede convertir fácilmente en un conjunto de reglas de clasificación. Los algoritmos de árboles de decisión que se utilizarán son REPTree, J48, RandomTree y ADTree.

En definitiva, es importante señalar que no se ha detectado un consenso entre los anteriores algoritmos de clasificación sobre la existencia de un único factor que más influya en el uso de TIC de los estudiantes. En cambio, si se pueden considerar el siguiente grupo de factores (que son los que más aparecen en los modelos obtenidos) como los más influyentes: a mayor conocimiento sobre el uso de TIC es mayor el grado académico obtenido, son hombres y tienen un promedio de edad de 25 a 35 años; el uso de TIC ayuda en la comprensión de un tema; las herramientas de

comunicación, tales como foros de discusión, intercambio de archivos, correo electrónico, anotaciones, chat, servicios de video, tablero o pizarra electrónica son aceptadas ampliamente; la estructura metodológica abre posibilidades de cursos no tan estructurados; el uso de TIC permite autoevaluarse y autocriticarse; las redes sociales son ampliamente utilizadas por jóvenes menores de 25 años.

Conclusiones y trabajo a futuro

Es importante mencionar que una tarea muy importante en este trabajo fue la recopilación y el pre-procesado de los datos, ya que la calidad y fiabilidad de la información afecta de manera directa en los resultados obtenidos. Es una tarea ardua, que implica invertir mucho tiempo y disposición.

Respecto a los resultados de clasificación de las diferentes pruebas, las principales conclusiones son:

- Se ha mostrado que los algoritmos de clasificación pueden utilizarse con éxito para identificar el uso académico de las TIC de los estudiantes.
- Se requiere de técnicas de selección de características cuando se dispone de muchos atributos, consiguiendo mejorar la clasificación de los algoritmos al utilizar un conjunto reducido.

Respecto del conocimiento extraído de los modelos de clasificación obtenidos, las principales conclusiones son:

- La utilización de algoritmos de clasificación de tipo “caja-blanca” permiten obtener modelos comprensibles por un usuario no experto en minería de datos en procesos de toma de

decisiones. En nuestro caso el objetivo final es poder detectar los usos de las TIC de los estudiantes universitarios de la modalidad a distancia para obtener elementos que nos permitan mejorar el diseño de las situaciones de aprendizaje en contextos formales a distancia y aprovechar la identificación y uso que estos muestran por determinados recursos tecnológicos.

Finalmente, como trabajo a futuro de esta investigación se van a realizar más pruebas utilizando otros algoritmos de minería de datos. Además de:

- desarrollar un algoritmo propio de clasificación/predicción para poder compararlo con los resultados de algoritmos clásicos y obtener mejores resultados de predicción.
- Intentar predecir el éxito o fracaso de los estudiantes ante situaciones de aprendizaje que emplean recursos tecnológicos.

Proponer métodos para ayudar a los estudiantes detectados dentro del grupo de riesgo. Posteriormente comprobar que porcentaje de las veces fue posible evitar que un estudiante detectado a tiempo fracasara o abandonara alguna situación de aprendizaje.

**Referencias**

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. Quaderns d'Estadística i Investigació Operativa, 25, num 3., 479-498.

Romero, C. and Ventura, S. (2007). Educational data mining: A Survey From 1995 to 2005. Expert System with Applications, 33, 135-146.

International Educational Data Mining Society. (s.f.). Recuperado el 29 de septiembre de 2013 de http://www.educationaldatamining.org/

Kotsiantis, S., Patriarcheas, K. and Xenos, (2010). A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´ Performance in Distance Education. Knowledge Based System, 23, no. 6, 529-535.

Kotsiantis, S. (2009). Educational Data Mining: A Case Study for Predicting Dropout – Prone Students. Int. J. Knowledge Engineering and Soft Data Paradigms, 1, no. 2, 101–111.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. and Loumos, V. (2009). Dropout Prediction in e-learning Courses through the Combination of Machine Learning Techniques. Computers & Education, 53, 950–965.

Weka. (s.f.). Recuperado el 1 de septiembre de 2013 de http://www.cs.waikato.ac.nz/~ml/weka/