APORTES ESPERADOS DE LA TÉCNICA DE ÁRBOLES DE DESICIÓN AL APLICARLOS A DATOS GENERADOS CON LA METODOLOGIA BLENDED LEARNING

Marcelo Omar Sosa, Sosa Bruchmann Eugenia Cecilia

Departamento Computación/Facultad de Ciencias Exactas y Naturales/Universidad Nacional de Catamarca
Av. Belgrano Nº 300 - Planta alta - C.P: 4700 - San Fernando del Valle de Catamarca
Teléfono: 0383- 4425610 /4420900
sosamod1@hotmail.com, sosab_ec@hotmail.com

Resumen

La búsqueda información oculta y útil que se encuentra en los datos de las actividades educativas es el objetivo principal de la minería de datos educativa o educational data mining (E.D.M. por sus siglas en inglés). En el caso particular de la aplicación de la metodología de blended learning en la educación superior, ésta genera grandes volúmenes de datos de dos tipos: los que se obtienen con la observación de las actividades presenciales y los que se almacenan con las actividades que desarrollan los alumnos en los entornos virtuales que se utilizan en esta metodología. Para completar el perfil del alumno a analizar, deben incorporarse además los datos históricos que los acompañan en el ingreso al nivel superior. Estos datos representan un gran volumen por lo que las técnicas de data mining representan un alternativa válida para su procesamiento. Dentro de ellas se encuentran los árboles de decisión que representan unas de las más importantes y utilizadas por los investigadores. El estudio del modo de aplicación y resultados obtenidos son la esencia de la presente investigación.

Palabras clave: Técnicas de data mining, árboles de decisión, blended learning.

Contexto

El presente trabajo se encuentra dentro de las líneas de investigación en que se desarrollan las actividades del equipo de investigación al cual pertenecen los autores. Actualmente los docentes están a cargo de asignaturas de las carreras del área informática de la Facultad de Ciencias Exactas y Naturales *(Fa.C.E.N.)* de la Universidad Nacional de Catamarca *(U.N.Ca.).* Estas asignaturas se dictan en la campus central en la Capital de la provincia como así también en las subsedes de la provincia de Tucumán.

Introducción

La metodología de blended learning se han generalizado en la mayoría de las instituciones del nivel superior. Los sistemas de aprendizaje que utilizan la web como una herramienta de intermediación pedagógica permiten el desarrollo de las actividades que propone esta metodología generando grandes cantidades de datos. Recientemente se ha incrementado el interés por extraer información que se encuentra dentro de estos tipos de datos con la aplicación de las técnicas de data mining. El procesamiento de datos educativos dio origen a lo que se denomina data mining educativo (o educational data mining EDM) [1], esta disciplina tiene como objetivo el descubrimiento de información a través de la aplicación de las técnicas a los datos relacionados con el desempeño pedagógico generados en estas plataformas sumados a los datos de las actividades presenciales realizadas por el alumno. Se desea analizar estos datos para extraer información por diversas razones como lo son: comprender la forma en que

aprenden los alumnos, descubrir la mejor forma de organizar los materiales y actividades de la asignatura, determinar cuáles atributos son más representativos para ser utilizados en las predicciones, desarrollar nuevas tipologías de estudiantes y ajustar las existentes, determinar la formación de grupos, descubrir patrones de comportamientos, modificar las estrategias pedagógicas, entre varias otras.
La aplicación de las técnicas de data mining permite superar las dificultades que presentan las metodologías estadísticas en cuanto al manejo de grandes volúmenes de datos como así también de numerosas variables en el proceso de análisis.
El análisis de de datos educativos puede hacerse desde distintas áreas de acuerdo a las diferentes técnicas que se apliquen, estas son:

Descripción: Como su nombre lo indica lo que busca es describir características de los datos para encontrar un modelo que adecuado a la información procesada. En el caso educativo permite caracterizar a los alumnos de acuerdo a su desempeño académico de los niveles anteriores [3].
Predicción: Se busca un modelo que permita predecir o estimar el valor resultante de acuerdo a los valores de los atributos que describen los datos procesados [4].

Segmentación: pretenden agrupar los datos procesados creando separaciones entre ellos que describan las características principales [5].

Cada técnica aplicada representa un método o un enfoque conceptual para extraer la información de los datos. Estas pueden aplicarse por medio de varios algoritmos que indican los pasos a seguir para cada técnica. Gracias a ellas pueden predecirse resultados generando modelos predictivos o encontrar relaciones que generan modelos descriptivos.

En el caso particular de la predicción, unas de las técnicas más utilizadas es la de árboles de decisión la que puede utilizarse para realizar predicciones o clasificación de los valores procesados. Representa la forma más generalizada de representación gráfica y la que las personas que no son técnicos o expertos en la materia comprenden con mayor facilidad.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que permite determinar la decisión final que se debe tomar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, entre otros [6].

La técnica de árbol de decisión puede ser implementada utilizando diversos algoritmos, entre los más utilizados podemos nombrar:

El algoritmo ID3 que fue desarrollados por John Ross Quinlan en el año 1986, su principal característica es que es resistente al ruido en los datos tratando de aproximar una función objetivo. Construye en árbol de decisión analizando la entropía, basándose en información y ganancia. Utiliza la estrategia denominada “divide y vencerás”.

El algoritmo 1R de Robert Holte, el cual utiliza un solo atributo para la clasificación. Tiene la ventaja de ser más simple que otros algoritmos y sus resultados pueden ser muy acertados en comparación con algoritmos más complejos.

El algoritmo J48 es una implementación del algoritmo C4.5, se basa en la incorporación del concepto de ratio de ganancia (gain ratio). Empareja las posibilidades de las variables impidiendo que aquellas variables con mayor número de valores posibles sean seleccionadas por encima de las demás. Es una evolución del algoritmo ID3 ya que permite trabajar con valores continuos y la poda del árbol una vez inducido.

El algoritmo 1R fue propuesto por Rober C. Holte. Únicamente utiliza un atributo para la clasificación. Selecciona un atributo del cual

sale una rama para cada valor que va a parar a la clase con más probabilidades. Cada rama del árbol generado está etiquetado como missing si los valores son desconocidos, un intervalo o un valor nominal dependiendo del tipo de variable.

El algoritmo PRISM supone que los datos no tienen ruido, se basa en la creación de reglas que maximicen la relación entre la cantidad de ejemplos positivos cubiertos y la cantidad de ejemplos cubiertos por la regla (Cant.posit./Cant.cub.) [2].
Se caracteriza por mostrar las reglas en forma ordenada ya que descarta los ejemplos que se van cubriendo con las reglas creadas en cada iteración.

Algoritmo PART trabaja generando las reglas sin optimizarlas para ahorrar en trabajo computacional. Su proceso consiste en generar las reglas para ir eliminando los ejemplos de entrenamiento que se van cubriendo, esto se repite hasta que no quedan ejemplos que procesar.

Existen otros algoritmos que se utilizan también en la creación de árboles de decisión como lo son: CLS (concept learning system) creado por Hoveland y Hunt en 1950, CHAID creado por G.V. kass en 1980, ID4 e ID4R J Shlimmer y D. Fisher en 1996, ID5 e ID5R P. Utgoff en 1990 y las mejoras sobre los algoritmos J48 y C5.0.

Metodología

El tema de investigación para a la realización del procesamiento de los datos requiere la selección de aquellos algoritmos que tengan mejor desempeño con respecto al conjunto de entrenamiento. Por ello para evitar errores de cálculo o de implementación de los algoritmos se plantea la utilización de un software que permita el cálculo utilizando varios algoritmos para realizar la comparación de los resultados. Si bien existen actualmente varios softwares que tienen la función planteada se selecciona WEKA por su gratuidad, simplicidad, cantidad de ejemplos y ayudas en línea. Entre sus ventajas se encuentran además la de poder presentar los resultados del procesamiento en forma gráfica mostrando el árbol de decisión resultante.

Para la generación de los arboles se utilizarán los datos recopilados de la asignatura Programación 1 correspondiente a las carreras de Profesorados en Computación y Tecnicatura en Informática de todas sus orientaciones de la Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de Catamarca.

Se crearan dos conjuntos de datos uno para desarrollar el entrenamiento en la construcción de los árboles y otro para la verificación de su funcionamiento.

Para organizar adecuadamente los resultados se prevé la generación de tablas comparativas en las que se vuelcan los resultados obtenidos en las diferentes corridas del programa, en cada una de ellas se seleccionará un algoritmo en particular. Esta tabla permite la selección del algoritmo con mejor desempeño en cuanto a la predicción de los resultados que obtendrán los alumnos a lo largo del curso de programación.

Conclusión

Se espera que la utilización de la técnica de árboles de decisión y su implementación con el algoritmo seleccionado permita la predicción del desempeño que presentarán los alumnos mientras recorren la asignatura Programación 1. Esta predicción permitirá la toma acciones sobre los alumnos en riesgo de abandono o desaprobación antes de que esto suceda.

Líneas de investigación y desarrollo

El trabajo se enmarca en la investigación de aplicación de técnicas y algoritmos de minería de datos en bases de datos educativas. Con el principal objetivo de cuantificar la efectividad del proceso de enseñanza y aprendizaje, organizar adecuadamente el contenido de la

asignatura, aplicar el agrupamiento de alumnos de acuerdo a los perfiles encontrados, realizar predicciones sobre el rendimiento de los alumnos, desarrollar actividades de apoyo a alumnos de riesgo de abandono entre otros.
El presente trabajo está íntimamente relacionado con trabajos desarrollados en el área y próximos a ser presentados en encuentros de investigadores en el tema. Pertenece al tema general que se desarrolla como tesis de posgrado de la carrera de maestría en ciencias de la computación que será presentada en la Universidad Nacional del Sur en el presente año.

Resultados y Objetivos

Los resultados esperados para este trabajo son los de analizar la información que proporciona la aplicación de esta importante técnica de data mining al ser aplicada a la base que contiene los datos obtenidos de las actividades desarrolladas por los alumnos utilizando la metodología de blended learning.

Tiene como objetivos principales:

- Comprender el funcionamiento de la técnica de data mining de arboles de decisión.

- Analizar la información que proporciona la aplicación de esta técnica a los datos producidos por la utilización de la metodología de blended learning..

- Determinar un método de selección de atributos que sean más representativos de las características de los alumnos que desarrollan las actividades con esta metodología.

- Buscar herramientas de datamining de uso libre que utilice la técnica de árboles de decisión para ser utilizada en la investigación.

Formación de Recursos Humanos

Los autores del trabajo se encuentran en la etapa de desarrollo de sus tesis de posgrado en carreras relacionadas con el tema de investigación, como la Maestría en Ciencias de la Computación en donde el Mgter. Sosa Marcelo Omar D. desarrolla actualmente la tesis "*Estudio y selección de técnicas y herramientas de data mining para ser aplicados a bases de datos utilizados en el blended learning*" que se realiza con la dirección del Dr. Chezñevar Iván de la Universidad Nacional del Sur (*U.N.S*). Además el investigador se encuentra en la etapa de planificación de su tesis doctoral en el área de minería de datos en el Doctorado en Ciencias dictado en la Facultad de Ciencias Exactas y Naturales (*Fa.C.E.N.*) en convenio con la Universidad Nacional del Sur (*U.N.S*).
La Docente Investigadora Lic. Eugenia Cecilia Sosa Bruchmann desarrolla su tesis en la carrera Especialización en Ingeniería en Software de la Universidad Nacional de San Luis denominada "La experiencia del usuario desde un nuevo enfoque para el desarrollo de productos interactivos: el comportamiento emocional del usuario y la importancia de los atributos estéticos" dirigida por el Dr. Germán Montejano. Los docentes dirigen tesis de la carrera de Licenciatura en Tecnología Educativa del alumno Albornoz, Marcelo denominada "Estudio del acceso al conocimiento de Computación de niños de las salas de cinco años del Nivel Inicial de los Jardines de Infante nuclearizados Nro. 7 de la zona norte de la Capital de Catamarca." la cual fue presentada y aprobada durante el año 2014.

Además desarrollan las siguientes actividades:

- Dirección de proyectos de investigación de voluntariado y pertenecientes a la facultad a la cual pertenecen.
- Integración de equipo de investigación de centro de investigación en Estadística de la Facultad de Ciencias Exactas y Naturales del U.N.Ca.
- Producción de artículos científicos para su presentación en congresos locales, nacionales e internacionales.

- Participación de los integrantes en cursos de actualización y posgrado en el área de estudio.
- Integrantes de la revista de ciencias de la Facultad de Ciencias Agrarias de la U.N.Ca.
- La actualización y capacitación permanente de los investigadores en talleres o workshop relacionadas con el tema del trabajo.
- La participación de los investigadores como consultores en proyecto afines que se desarrollan en la Facultad de Ciencias exactas y Naturales en distintas áreas.
- Examinadores de trabajos finales en las diferentes carreras que se dictan en la Fa.C.E.N. de las U.N.Ca.
- Dirección de tesis y tesinas finales de las carreras de computación, informática y Licenciatura en tecnología educativa.
- La planificación de seminarios para docentes en temas relacionados con la investigación y resultados obtenidos en la investigación.
- Participación en convenios con la Facultad de Tecnología para el desarrollo de estudios del área de datamining.

REFERENCIAS

1. *C. Romero and S. Ventura, "Educational data mining: A Survey From 1995 to 2005", Expert System with Applications, vol. 33, pp. 135-146, 2007.*

2. *Lavrac, N., Kavsec, B., Flach, P todorovski, L., "Subgroup discovery with CN2-SD". Jornal of machine learning research. 2004.*

3. *Jain A.K.and Dubes R.C."Algorithsm for clustering data. 1998. Englewood Cliffs. N.J. Prentice Hall.*

4. *Fayyad, U, Piatetsky-Shapiro, G. and Smyth, P., "The KDD process for extracting usefull knowledge from volumes of data". Communication of ADM 1996.*

5. *Agrawal, R., Imiellinsky, T. and Swami, A.N.."Mining Assosiation Rules between set of item in large databases". In International conference on management of data. 1993. Whashington D.C. ACM Press.*

6. *Solarte Martinez, Guillermo Roberto, Ocampo S., Carlos Alberto. Técnicas De Clasificación Y Análisis De Representacion Del Conocimiento Para Problemas De Diagnósticoscientia Et Technica [En Linea] 2009, Xv (Agosto-Sin Mes) : [Fecha De Consulta: 16 De Marzo De 2015].*

BIBLIOGRAFÍA Y TRABAJOS CONSULTADOS

- *Gabriel Páramo and Carlos Correa, "Deserción Estudiantil Universitaria. Conceptualización," Medellín, Revista Abril - Mayo – Junio 1999.*

- *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, pp. 37-54, 1991.*

- *Daniel T Larose, Discovering Knowledge in Data: An Introduction to Data Mining. New York: John. Wiley & Sons, 2004.*

- *Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students' Performance," in International Journal of Advanced Computer Science and Applications, India, 2011, pp. Vol. 2, No. 6.*

- *Baker Ryan and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," JEDM - Journal of Educational Data Mining, vol. 1, no. 1, Octubre 2009.*

- *Jonathan E. Freyberger, Neil He®ernan,*

and Carolina Ruiz. Using association rules to guide a search for best ¯tting transfer models of student learning. Master's thesis, Worcester Polytechnic Institute, 2004.

- *Merceron and K. Yacef. Educational data mining: a case study. Process of 12th. Conference on Artificial Intelligence in Education (AIED03), page 467 .2005.*

- *María Delia Grossi.”Reglas de predicción aplicables al diseño de un curso de computación”. Marzo 2008.*

- *Erwin Sergio Fischer Angulo.”Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios”. Santiago de Chile 2012.*

- *Pedro Gonzalez Garcia,.”Aprendizaje evolutivo de reglas difusas para la descripción de subgrupos”. Granada España. 2007.*

- *Cristoban Romero, Sebastian Ventura, Nykola Pechenizkiy and Rayan Beker.”Handbook of educational data mining”.Chapman & Hall CRC press. 2011.*