Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo

Alejandro Ballesteros Román, **Daniel Sánchez-Guzmán and Ricardo García Salcedo**
Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Unidad Legaria del Instituto Politécnico Nacional. Calzada Legaria, No. 694, Col. Irrigación. Del. Miguel Hidalgo, C. P. 11500, México D. F. México.

E-mail: baroal87@gmail.com

(Recibido el 25 de Junio de 2013, aceptado el 14 de Noviembre de 2013)

Resumen

En la actualidad se han desarrollado e incorporando diferentes estrategias de aprendizaje para la enseñanza en disciplinas que por su naturaleza se consideran ciencias duras como puede ser Matemáticas, Física y Química; de igual manera pero en menor cantidad se han implementado estrategias de aprendizaje en Ingeniería y Ciencias Computacionales. Considerando que estas estrategias deben de crecer en cantidad y calidad, aparece un problema inherente que tiene que ver con el diseño instruccional. Este problema es la medición de la efectividad tanto de manera cualitativa como cuantitativa del avance en el aprendizaje por parte de los estudiantes. Al realizar la implementación de un diseño instruccional o un experimento educativo, este genera una cantidad considerable de información derivada de las diferentes actividades llevadas a cabo por los estudiantes, como pueden ser los instrumentos de medición, los reportes escritos, las exposiciones de los estudiantes, entre otras actividades; por su naturaleza se puede observar que cada actividad mencionada anteriormente genera una cantidad de información a través de las evidencias proporcionadas lo que eleva esta cantidad de información. Si los datos son considerablemente grandes se presenta un problema de intratabilidad de la información y posiblemente no se puedan analizar otras variables que pueden ser importantes para su estudio. El presente trabajo presenta un estudio de los métodos y técnicas de Minería de Datos Educativa (Educational Data Mining – EDM, por sus siglas en inglés); aplicada a la educación, como una herramienta donde el investigador podrá tener un margen mayor de análisis de la información a través de los datos generados por la aplicación, apoyándose de igual manera en algoritmos de inteligencia artificial.

Palabras clave: Minería de Datos Educativa, Patrones de Comportamiento, Tecnologías de la Información y Comunicación, Agentes Inteligentes, Granularidad Educativa.

Abstract

It´s currently developed and incorporating different learning strategies for teaching in disciplines which by their nature are considered hard sciences such as mathematics, physics and chemistry, same way but fewer have implemented learning strategies and Engineering Computer Science. Considering that these strategies need to grow in quantity and quality, shows an inherent problem that has to do with instructional design. This problem is the measurement of the effectiveness of both a qualitative and quantitative progress in learning by students. When implementing an instructional design or educational experiment, this generates a considerable amount of information derived from the various activities carried out by students, such as measuring instruments, written reports, presentations by students, among other things, by their nature can be seen that each above-mentioned activity generates a lot of information through evidence provided bringing this amount of information. If the data are considerably large is a problem of intractability of information and may not be able to analyze other variables that may be important for study. This paper presents a study of the methods and techniques of Educational Data Mining applied to education, as a tool where the researcher may have higher margin analysis information through the data generated by the application, based equally on artificial intelligence algorithms.

Keywords: Educational Data Mining, Behavioral Patterns, Technologies of Information and Communication, Intelligent Agents, Educational Granularity.

PACS: 01.40.-d, 01.40.Fk, 01.50.-I, 01.50.Kw, 07.05.Mh **ISSN 1870-9095**

I. INTRODUCCIÓN

Las Tecnologías de Información y Comunicación (TIC) actualmente son temas de amplia investigación en los diversos sectores de investigación tecnológico por su constante evolución y aplicación dentro de los diferentes procesos y actividades del ser humano, de tal forma que se tiene la integración de servicios y aplicaciones para los

usuarios finales con la finalidad de facilitar y hacer accesible el manejo de la información con el objetivo de solucionar los problemas cotidianos, como por ejemplo; sistemas bancarios, sistemas educativos, sistemas de posicionamiento global (GPS), por mencionar algunos, de manera que todo esto ha permitido desarrollar lo que actualmente se conoce como sociedad de la información. Con el transcurso de los años, las actividades y el desarrollo de nuevas tecnologías han generado de forma considerable el almacenamiento de información, donde todo ese flujo de información que sea recolectado ha permitido satisfacer las necesidades diarias de las organizaciones, pero ha presentado un problema inherente en las capacidades humanas para analizar y transformar la información en conocimiento útil y relevante que apoye a la toma de decisiones. Este tipo de problemas comúnmente se han presentado en los grandes sistemas conocidos como bases de datos, en los cuales almacenan una gran cantidad de información; estos datos con frecuencia contienen valioso contenido y pueden ser vistos como una recopilación masiva de información para el uso de la mayoría de las organizaciones, formando parte importante en el área de conocimiento informático, aun con el uso de herramientas estadísticas clásicas para la manipulación, el análisis y la extracción de alguna información importante está tarea es casi imposible, posteriormente esta necesidad de intratabilidad ha motivado al empleo de técnicas y herramientas de Minería de Datos (Por sus siglas en ingles DM – Data Mining), que posibiliten la extracción de conocimiento en forma de reglas y patrones a partir de dichos datos, es decir, es un proceso de análisis de datos que consiste en buscar o encontrar tendencias o variaciones de comportamiento en los datos denominados patrones. De tal forma, hoy en día este tipo de problemas se ha hecho inherente en el sector educativo tanto público como privado, varias disciplinas como las ciencias de la computación, matemáticas y educación se han enfocado en las técnicas de la DM para realizar un estudio y análisis más detallado y preciso sobre el comportamiento que tienen los alumnos en los diversos sistemas de aprendizaje-enseñanza, en el cual generan una gran cantidad de información por ejemplo un sistema virtual que se apoya en el uso de agentes tutores inteligentes [1] en el cual los estudiantes interactúan y aprenden un tema en particular, estas transacciones son almacenadas en las bases de datos y las mismas se componen de información, como puede ser: el tiempo en resolver un problema, oportunidades permitidas a los estudiantes, sí la respuesta fue acertada o equivocada, etc. Derivado de lo anterior se puede identificar que no se analiza el comportamiento que tiene el estudiante dentro el modelo educativo que se está implementando. El objetivo de este artículo es presentar un estudio de una herramienta que implemente las técnicas y métodos de lo que se ha denominado actualmente como Minería de Datos - DM enfocada a la gestión educativa, esta herramienta servirá a docentes que requieran evaluar sus prácticas y metodologías desarrolladas e implementadas con sus estudiantes, ya sea mediante el uso de agentes tutores inteligentes, sistemas virtuales de educación o estrategias activas de aprendizaje dentro del aula. Su objetivo principal: es procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil para un usuario y satisfacer sus metas respecto al conocimiento a detalle de un sistema de información sobre un contexto educativo.

II. DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

En esta sección se describirá uno de los métodos analíticos para la selección de atributos y la extracción de parámetros de información, para lograrlo es necesario preparar correctamente los datos para procesarlos, elegir un método adecuado para la extracción de los patrones deseados y finalmente determinar cómo evaluar los patrones encontrados, estas etapas han sido organizadas en un esquema conocido como el proceso de Descubrimiento de Conocimiento en Bases de Datos (por sus siglas en ingles KDD – Knowledge Data Discovery) donde una de las principales etapas utiliza la DM, asimismo se presenta una breve explicación del proceso KDD, describiendo cada una de las etapas.

A. Knowledge Data Discovery

El proceso KDD es un proceso no-trivial de identificación de patrones válidos, novedosos y potencialmente útiles sobre un conjunto de datos, esto es, el objetivo es encontrar conocimiento útil, valido, relevante y nuevo sobre una determinada actividad. En este contexto los datos hacen referencia a un conjunto de hechos o ejemplos en una base de datos y los patrones son resultados o expresiones en algún lenguaje que puedan describir de manera compacta los datos asimismo el termino no-trivial se comprende qué alguna búsqueda o inferencia es llevada a cabo, esto es, implica la búsqueda de modelos, estructuras, patrones o parámetros [2], [3]. De manera que el proceso KDD está dividido por una serie de etapas en el cual se estructura por tres grandes bloques los cuales son: el pre-procesamiento, búsqueda o identificación de patrones (La utilización de técnicas y métodos de DM) y la evaluación, esto es, está dividido por una serie de pasos desde la selección y limpieza de la base de datos hasta la evaluación e interpretación de los resultados tal como se ilustra en la Figura 1.

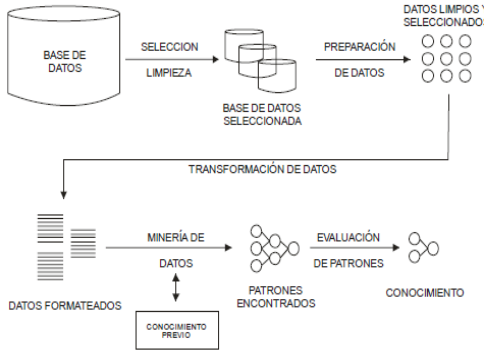

FIGURA 1. Proceso esquemático de la Minería de Datos aplicada a una base de datos en general.

Por lo tanto, de la ilustración anterior se puede observar que el proceso KDD se estructura por una serie de pasos iniciando por la selección, preparación, limpieza y el formateo de los datos de acuerdo a los patrones analizar, a esta etapa es conocida como el *pre-procesamiento*, posteriormente interviene la etapa de la *minería de datos* en el cual tiene como tarea buscar y descubrir patrones ocultos en las bases de datos en base a la utilización de algún algoritmo de DM a implementar, pasando a la última etapa de *evaluación*, determinando la validez y confiabilidad del conocimiento adquirido, es decir, los patrones deben ser válidos y de alto impacto para el usuario final [2, 3]. A continuación se describen las etapas del proceso KDD para enmarcar el objetivo general del funcionamiento de este tipo de sistemas.

B. Pre-procesamiento

Dentro de esta etapa se determina la preparación de los datos para implementar la siguiente etapa donde se implementaran las técnicas de la minería de datos, por consiguiente el pre-procesamiento se rige por tres pasos básicos, los cuales son la selección, limpieza, preparación y transformación de los datos. Selección y Limpieza de los Datos; este proceso se encarga de determinar las fuentes y características de la información, esto es, permite la navegación y visualización previa de los datos determinando que aspectos son de interés y puedan ser estudiados asimismo existen varias bases de datos que tienen diversas inconsistencias de tal manera que la limpieza y el procesamiento de datos involucra una estrategia para manejar adecuadamente el ruido que contengan algunos datos los cuales pueden ser valores faltantes, inconsistencias en los valores que no corresponden a los dominios de los atributos o que puedan ser contradictorios, esto es, valores incompletos o erróneos en la fuente de información por mencionar algunos. De tal forma que este tipo de problemas deben eliminarse antes que la etapa de DM, de tal forma que puedan afectar la precisión de los resultados o incluso el algoritmo, puede construir modelos ineficientes a partir de un conjunto de datos incorrectos. Preparación de los Datos; Este proceso busca eliminar los datos que no serán relevantes para el procesamiento de la etapa de la minería de datos, no todas las bases de datos necesitaran la aplicación de todas las etapas del pre-procesamiento, por ejemplo una base de datos que tiene los registros de los estudiantes en el cual almacena el estatus de un curso, si todos los atributos son significativos después de eliminar las inconsistencias el proceso omitirá dicha etapa pasando a la transformación de los datos, la tarea consiste en identificar características específicas de los estudiantes en la Transformación de Datos; este proceso consiste en la transformación de los datos, donde cada algoritmo que se implementara siempre establece el tipo y estructura de los datos con los que procesara, esto es, cada algoritmo de minería de datos a utilizar requieren un formato y la estructura para sus entradas, de tal forma que la tarea que se está resolviendo los datos no tienen la entrada establecida por el algoritmo entonces se procederá a transformarlos [2, 4, 5].

C. Minería de Datos (DM)

Esta etapa es considerada como la parte central del proceso KDD, con la finalidad de encontrar o descubrir los patrones de interés para el usuario final, estos patrones pueden ser grafos, reglas de asociación, clasificaciones, una red neuronal, clustering, entre otros. La tarea que realiza la DM son: Seleccionar y aplicar el método de DM apropiado, es decir, la realización de una selección de la tarea para el descubrimiento del conocimiento, tales métodos son como la clasificación, agrupamiento (Clustering), reglas de asociación, regresión, por mencionar algunas. Este sección lleva a cabo el proceso de la DM en busca de patrones para expresarlos en modelos o simplemente la expresión de dependencias de datos, este modelo depende de su función (clasificación) y de los métodos de representación por la elección de algún algoritmo como los Arboles de Decisión, Reglas de Asociación, el teorema Naive – Bayes, algunos métodos de Inteligencia Artificial como las Redes Neuronales, etc., asimismo se tiene que especificar un criterio de preferencia para la selección de un modelo dentro de un conjunto posible de modelos, también es necesario especificar la estrategia de búsqueda a utilizar que normalmente se encuentra dentro de algún algoritmo de DM [2, 4, 5].

D. Evaluación

Esta etapa se considera como la evaluación, interpretación, transformación y representación de los patrones extraídos del proceso KDD donde se establecen parámetros que permitan comparar la calidad de los resultados obtenidos y su validación por medio de modelos representativos de manera gráfica como curvas de aprendizaje, tasas de error, perfiles de rendimiento por mencionar algunos. Esto es, el análisis de los resultados obtenidos sobre el proceso de la DM y posiblemente repetir los pasos anteriores, es decir, repetir el proceso si los resultados obtenidos no fueron satisfactorios para un modelo cualitativo o se requiera implementar un nuevo algoritmo o quizá generar un análisis de nuevos datos y la implementación de nuevas estrategias que sean de utilidad para el usuario final. Asimismo esta etapa final del proceso KDD implica que el conocimiento obtenido realice las acciones requeridas para el buen desempeño del sistema o para almacenarlo y reportarlo a los usuarios interesados [2, 4, 5].

III. MINERIA DE DATOS

En esta sección se presenta una descripción general sobre el estado actual de la Minería de Datos (Por sus siglas en ingles DM - Data Mining) para dar paso al término de Minería de Datos Educativa (Por sus siglas en ingles EDM – Educational Data Mining), mostrando todos los componentes que son necesarios para el desarrollo e implementación de los métodos y técnicas para la interpretación y el análisis de patrones de conocimiento, asimismo las diferentes teorías

que conllevan a la aplicación de un método estadístico, sus implementaciones y la selección de los diversos métodos para ser empleadas para un análisis proactivo y eficiente. La DM es considerada como un mecanismo de explotación que consiste en la búsqueda y el descubrimiento de información valiosa dentro de grandes volúmenes de datos (Bases de datos), esto es, un proceso de análisis que trabaja a un nivel de conocimiento con el fin de descubrir patrones, relaciones o incluso excepciones útiles asimismo la generalización de modelos predictivos que proporcionen patrones de conocimiento de alto impacto y puedan llegar a la toma de decisiones, la DM utiliza gran cantidad de métodos como la estadística, inteligencia artificial, computación gráfica, procesamiento masivo de conjuntos de información y como materia prima las bases de datos, [9]. El desarrollo de la DM es una actividad profesional, indispensable para distinguir previas actividades de modelos estadísticos y de amplias actividades de descubrimiento de conocimiento que se generan constantemente en los sistemas de almacenamiento, en base a las siguientes consideraciones teóricas; El Descubrimiento de Conocimiento considerado como el proceso de acceso a los datos, es decir, la exploración, preparación, métodos de implementación (en base a los algoritmos de DM), modelado y la supervisión de las actividades o métodos de la DM. Modelos Estadísticos (utilización de métodos y técnicas de la DM, tal y como se puede observar en la Figura 2) en base a la implementación de algoritmos estadísticos paramétricos para la agrupación o predicción de resultados o eventos en base a la representación de patrones o variantes de conocimiento interesante, [7, 9]. De manera que una definición de la DM de acuerdo a Usama Fayyad, 1996; "Es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos".

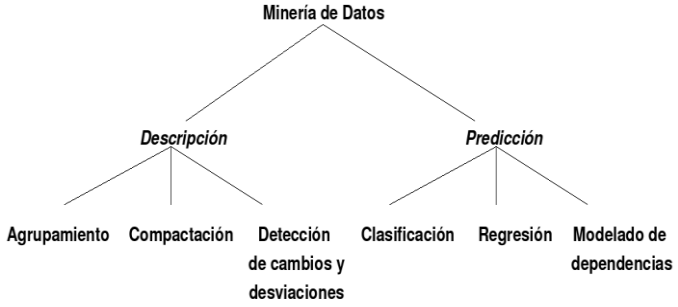

FIGURA 2. Clasificación y relación entre los diferentes algoritmos de Minería de Datos.

En el ámbito de los negocios la DM ha atraído la atención de las industrias de información y la sociedad en su conjunto sobre el paso de los años, en consecuencia a la amplia disponibilidad de grandes cantidades de datos y como resultado ha permitido convertir esos datos en información útil y conocimiento para las grandes y pequeñas organizaciones, [9]. Esta información de conocimientos adquiridos son utilizados para la aplicación que va de la retención de análisis de información, detección de fraudes, control de la producción y la exploración en la ciencia, la industria en donde los sistemas de bases de datos han tenido un gran camino evolutivo en el desarrollo de la recopilación y almacenamiento de datos y la creación de las mismas, así como el análisis avanzado. De tal forma que la meta u objetivo es ayudar a buscar situaciones interesantes con los criterios correctos, complementar una labor que hasta ahora se ha considerado "intelectual" y de alto nivel, privativa de los gerentes, planificadores y administradores, además de realizar la búsqueda usando tiempos de máquina excedentes, [4].

E. La Minería de Datos en el Sector Educativo

En el sector educativo, las técnicas de minería de datos son usadas para la comprensión del comportamiento de los estudiantes, la MDE emerge como un paradigma orientado para la generalización de modelos, tareas, métodos y algoritmos para la exploración de datos que provienen de un contexto educativo asimismo tiene como función encontrar, analizar patrones que caractericen los comportamiento en base a sus logros, evaluaciones y el dominio de contenido de conocimiento que tienen los alumnos en los diversos mecanismos de aprendizaje-enseñanza que hoy en día son otorgados en las diversas instituciones públicas y privadas con el objetivo de generar modelos educativos en los cuales puedan fomentar nuevas técnicas o herramientas que puedan analizar e incrementar el nivel participativo de los estudiantes sobre los sistemas de aprendizaje-enseñanza. Por ejemplo la recomendación de actividades para ofrecer nuevas experiencias de aprendizaje, avisos o predicciones de rendimiento de los alumnos para mejorar la efectividad del curso o promover el trabajo en grupo, por mencionar algunos, [6, 12]. Todas estas modalidades generan información de manera directa e indirecta, ya sea por las interacciones del estudiante con sus compañeros, con el docente y con las herramientas tecnológicas que se encuentran a su disposición para poder interactuar y recibir la instrucción correspondiente, dichos datos provienen de varias fuentes de información principalmente en las aulas donde el docente y el estudiante intercambian información en el cual desarrollan y aplican estrategias de aprendizaje sobre un medio de apoyo (la utilización de las Tecnologías de Información y Comunicación – TIC), [6]. Ahora bien, no sólo existe información de manera presencial, en el ámbito educativo existen diferentes modalidades de aprendizaje, teniendo presente al menos tres bien definidas; la primer modalidad es la presencial que es cuando los estudiantes acuden a las escuelas para asistir a clases y los docentes exponen los temas de diferentes maneras, la segunda modalidad es la mixta, dónde los estudiantes acuden un cierto periodo a la escuela para recibir instrucción y en otro periodo tienen actividades en línea o virtual a través del uso de Internet (Sistemas E-learning); por último se tiene la modalidad virtual o en línea, (la utilización de Sistemas de Agentes Tutores Inteligentes, [1]) donde tanto el estudiante como el docente interactúan a través de Internet y mediante plataformas de intercambio de información, en esta modalidad nada es realizado de manera presencial, [1, 6]. Por lo tanto, el sector educativo enfrenta un gran reto hoy en día sobre todo relacionado con el pronóstico de las

trayectorias que generan los estudiantes en la iteración que estos tienen con los sistemas educativos aprendizaje-enseñanza; ya que las instituciones educativas generan constantemente gran cantidad de información por lo que les interesa saber que está pasando con ese flujo de información, es decir, que acciones provocan que los estudiantes se interesen en el estudio de ciencias duras como las matemáticas, la física, la química, ciencias computacionales, humanidades o artes, ya que son variadas las estrategias educativas aplicadas, así como saber cuáles podrían ser los perfiles de ingreso y egreso que mejor se adecuan a su modelo educativo, asimismo los posibles problemas de deserción o perdida de interés de los alumnos [6, 11]. Todo este entorno se ha enfocado a un nuevo término conocido como EDM, una disciplina emergente considerada como una alternativa tecnológica para alcanzar otras áreas que no se consideraban en los sistemas educativos, basada en los desarrollos de métodos y técnicas para el análisis y la exploración de ciertos tipos particulares de datos obtenidos desde un contexto educativo, propiciando un alto impacto en los sistemas de aprendizaje-enseñanza. Por lo tanto, desde un punto de vista general la EDM conlleva una evaluación de un programa curricular o una unidad de aprendizaje que tiene como propósito incidir en el estudiante donde el instructor/investigador adquiera un conocimiento y lo convierta en aprendizaje y el alumno considerado como el usuario final se apropie del mismo llevándolo a un contexto de su vida cotidiana, [10]. De lo anterior se considera que la EDM nos conlleva a dos puntos de vista u orientaciones distintas; Orientado a los instructores/investigadores, con el principal objetivo de ayudar o apoyar a los educadores para mejorar el funcionamiento y el rendimiento de los sistemas de aprendizaje-enseñanza a partir del conocimiento adquirido sobre el flujo de información que se ha derivado del alumnado en base a modelos predictivos que puedan identificar de forma cualitativa y cuantitativa. Orientado a los estudiantes, con el principal objetivo de ayudarlo en la interacción con los sistemas aprendizaje-enseñanza incrementando sus experiencias, la implementación de nuevas herramientas para facilitar su conocimiento en los diversos temas educativos, sugerencias o actividades en el curso de acuerdo al progreso de aprendizaje, etc., [8]. El desarrollo de las técnicas de la EDM puede darse a partir de modelos supervisados o no-supervisados, esto es, la minería de datos supervisada; (aprendizaje a partir de ejemplos, con profesor) consiste en utilizar registros de los resultados que se conocen, por ejemplo, una base de datos de graduaciones que contienen registros de alumnos que han finalizado sus estudios y de los que aún siguen inscritos, esto lleva a vincular los patrones de conducta a los historiales académicos u otra información registrada, de manera que los ejemplos de entrada van acompañados por una clase o salida correcta, esta técnica se engloba al aprendizaje memorístico (Rote Learning), a los modelos de aprendizaje por ajuste de parámetros y a una amplia gama de métodos de construcción de varios modelos de clasificación, [11]. La minería de datos no-supervisada; (aprendizaje por observación) consiste en situaciones en las cuales se desconocen los patrones o agrupaciones en particular, por ejemplo, las bases de datos por curso de los estudiantes, esto genera poca información sobre los cursos que se realizan en un grupo o el tipo de curso que se relaciona con qué tipo de estudiante, por lo que la minería de datos no-supervisada se refiere al estudio y la búsqueda de patrones ocultos para conocer, clasificar y codificar el conocimiento antes de aplicar teorías, de tal forma que se construyen descripciones, hipótesis o teorías a partir de un conjunto de hechos u observaciones sin que se aplique una clasificación *a priori*, cabe destacar que este tipo de técnica o de aprendizaje son utilizados por los métodos de agrupamiento (Clustering), reglas de asociación, análisis de secuencia, [11]. De esta manera se define a la EDM partiendo de la principales teorías de la MD y complementada con las definiciones sobre el impacto en los sistemas aprendizaje-enseñanza, se puede proponer lo siguiente: "La Minería de Datos Educativa - EDM es el empleo de las herramientas tecnológicas, algoritmos y las estrategias de análisis de información utilizadas por la Minaría de Datos - DM, pero dentro de un contexto educativo para la búsqueda, análisis y la extracción de patrones de conocimiento, dónde se resuelvan problemas que mejoren el proceso enseñanza – aprendizaje a partir de modelos predictivos de forma cualitativa y cuantitativa". De tal forma que los resultados de la EDM pueden servir para que los investigadores, docentes y directivos puedan definir políticas de operación, adecuación al diseño instruccional vigente dentro de una institución, [6].

IV. METODOS Y TECNICAS DE LA MINERIA DE DATOS EDUCATIVA

Los clasificadores, agrupamiento y reglas de asociación son componentes de software fundamentales para la operación de la EDM, estos permiten identificar la información oculta para los diferentes actores dentro de las instituciones educativas. En esta sección se presenta una descripción general de los principales métodos y técnicas que son utilizados en la EDM, se describirá tanto el análisis, su construcción y el método para la generación de un modelo predictivo.

F. Clasificadores para Minería de Datos Educativa

La idea de la clasificación es colocar un objeto o categoría en base de sus otras características, para la creación de modelos predictivos que describan clases de datos importantes o predicciones de tendencias futuras tomando en cuenta que la DM busca obtener reglas para la división de los datos en una clase o categoría para la mejor comprensión de los datos, [8, 9]. La meta u objetivo de la clasificación es aprender o enseñar a una maquina la forma de clasificar los objetos a través del análisis de una hipótesis definida donde cuyas clases se conocen asimismo analizar el funcionamiento y el comportamiento de implementar los algoritmos de clasificación sobre un conjunto de ejemplos univariados y multivariados. De manera que las técnicas de clasificación se llevan a cabo de forma diferente según el método utilizado, sin embargo todas parten del mismo conjunto de datos considerado como muestra y todas terminan descubriendo

una predicción de acuerdo aún porcentaje de acierto, [8]. Por lo cual la forma básica de los clasificadores son llamados discriminativos por que determina un valor de la clase por cada registro del dato, es decir, si *M* es un clasificador, *C = {c1, c2,....., c}* el conjunto de valores de la clase y "*t"* un registro del dato (una tupla), entonces la predicción de la clase es *M(t) = ci* para un solo "*i"* (donde *i* es el valor del atributo), otra alternativa es un clasificador probabilístico, definiendo la probabilidad de una clase para todas las filas clasificadas, la predicción de la clase es *M(t) = [P(C = c1|t), … P(C = ci|t)]*, donde *P(C = ci|t)* es la probabilidad de que "*t"* pertenece a la clase c_i, [8]. La combinación de los clasificadores descritos anteriormente se puede obtener mejores resultados que de la utilización de los clasificadores de forma individual, por lo que estos enfoques son comparados para la determinación, su idoneidad y pertinencia de uso para la clasificación de datos típicos.

En el sector educativo, las técnicas de EMD en base a la clasificación y predicción son usadas para la comprensión del comportamiento de los estudiantes, de tal forma que pretende analizar el comportamiento que tienen los alumnos en los diversos mecanismos de aprendizaje que hoy en día son otorgados en los diversos sistemas de aprendizaje-enseñanza con el objetivo de generar modelos educativos que puedan analizar e incrementar el nivel participativo y de aprendizaje de los estudiantes en función de modelos predictivos que conlleven a un alto impacto y puedan ser implementados en la actualidad como pueden ser el modelo presencial, modelo virtual y/o modelo mixto, [6]. Asimismo en la educación, los instructores siempre están clasificando a los estudiantes según el comportamiento, su motivación y su conocimiento. De esta manera, el proceso enseñanza-aprendizaje, los sistemas de control escolar así como los resultados de implementar nuevas metodologías educativas como son los tutores inteligentes, video conferencias, software educativos por mencionar algunos, generan una cantidad enorme de información, la cual nos lleva a la aplicación de múltiples métodos de clasificación estudiados en la literatura que ofrecen resultados favorables en su combinación entre tiempo de ejecución y porcentaje de acierto, [8, 9, 10]. Sin embargo, antes de que el sistema pueda seleccionar una medida de adaptación como la selección de tareas, material didáctico, o método de aprendizaje, primero se debe clasificar el estatus actual del alumno, de tal forma; en primer lugar, tenemos que elegir el método de clasificación (Test set), que posteriormente serían los árboles de decisión, redes bayesianas, o redes neuronales, el clasificador del *k*-vecino más cercano, máquinas de vectores de soporte que son los más utilizados en la EDM. Posteriormente, necesitamos una muestra de datos (Training set), donde todos los valores de la clase se conocen, los datos se dividen en dos partes, un conjunto de entrenamiento y un conjunto de pruebas. El conjunto de entrenamiento se da a un algoritmo de aprendizaje, que se deriva de un clasificador, por consiguiente el clasificador se prueba con el equipo de prueba, donde todos los valores de clase se ocultan. Por lo tanto, estos métodos se aplican en clases de un conjunto de entrenamiento el cual genera gran cantidad de información valiosa, de manera que se pueden utilizar varios algoritmos para el descubrimiento de conocimiento, es decir, descubrir, encontrar o buscar la forma en que el vector de atributos de los casos se comporta y poder estimar las clases para nuevas instancias, [8].

G. Clustering para la Minería de Datos Educativa

Los métodos de agrupamiento (Clustering) a diferencia de la clasificación y la predicción que utilizan modelos supervisados analizan los registros con etiquetas de clase o categoría los objetos de datos (sobre la base de la clase de información de las etiquetas disponibles en conjuntos de prueba), de tal forma que los métodos de agrupamiento no están interesados en modelar un conjunto de relaciones y un conjunto de respuestas que pertenezcan a cada valor asociado, este método utiliza el modelado no-supervisado donde analiza los objetos de datos sin consultar con una clase conocida, es decir, las etiquetas de la clase no se encuentran en los datos de entrenamiento por el hecho de que el objeto es desconocido (Se ignora la similitud de los patrones), de tal forma que el clustering es utilizado para la generalización de etiquetas, esto es, la búsqueda y construcción de grupos formados de tal forma que los objetos (Patrones) dentro de un conjunto (un grupo) tienen una alta similitud en comparación entre sí denominada función de medida de asociación (similitud matemática en espacios métricos definida por medio de una norma de distancia), donde cada grupo formado puede ser interpretado como una clase de objetos de los cuales se pueden derivar reglas. Por lo tanto en una forma simplificada puede ser definida como la división o partición de un conjunto de observaciones con el objetivo principal de asignar a cada uno de los *N* objetos de datos a uno de los *K* posibles grupos disjuntos mediante una medida de similitud. Cabe mencionar que la similitud es un concepto difícil de clustering teniendo en cuenta que más allá de la densidad de los datos debemos considerar la forma y el tamaño del cluster. De manera que el objetivo principal del clustering es descubrir y modelar los grupos en que los elementos de datos se agruparan, es decir, encontrar los puntos de datos que naturalmente se agrupan dividiendo el conjunto de datos en categorías más comunes.

V. CONCLUSIONES

La Minería de Datos Educativa - EDM es un área multidisciplinaria en la cual convergen distintos paradigmas de computación como son el desarrollo o construcción de algoritmos de predicción, programación lógica, algoritmos estadísticos, entre otros, con el objetivo de generar principales tareas como; la clasificación, agrupamiento (Clustering), estimación, modelado de dependencias, visualización y descubrimiento de reglas con el fin de construir un modelo ajustado a un conjunto de datos sobre un contexto educativo, teniendo como fin último el proporcionar un conocimiento certero del sistema y predecir comportamientos futuros, asimismo el mejoramiento de los sistemas de aprendizaje-enseñanza. Por lo tanto de las teorías presentadas podemos decir que un patrón va a representar un conocimiento si excede algún umbral de interés en este caso

en el dominio educativo el conocimiento es especifico del dominio, es decir, se puede obtener un conocimiento profundo y comprensible con la finalidad de analizar el comportamiento y el sesgo productivo del alumno con el objetivo de poder resolver las hipótesis que se puedan generar así también está determinado por las funciones y umbrales sobre las interacciones de los sistemas aprendizaje-enseñanza.

AGRADECIMIENTOS

Alejandro Ballesteros Román agradece al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca otorgada para sus estudios de doctorado y al Programa Institucional de Formación de Investigadores (PIFI) del Instituto Politécnico Nacional (IPN) además del apoyo económico otorgado a través del proyecto SIP 2013-0169. Daniel Sánchez Guzmán y Ricardo García Salcedo agradecen al CONACyT por la beca como miembro del Sistema Nacional de Investigadores y al IPN por los recursos aportados a través de los proyectos SIP 2013-0169 y 2013-1811, a la COFAA-IPN por la beca otorgada y a la SIP-IPN por los apoyos mediante beca EDI.

REFERENCIAS

[1] Sánchez Guzmán, D., *Agentes Inteligentes; Diseño e Implementación para la Enseñanza de la Física*, Tesis Doctoral en Tecnología Avanzada, Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional, México (2009).

[2] Mondragón Becerra, R., *Exploraciones sobre el Soporte Multi-Agente BDI en el Proceso de Descubrimiento de Conocimiento en Bases de Datos*, Tesis de Maestría en Inteligencia Artificial. Departamento de Inteligencia Artificial, Universidad Veracruzana, México (2007).

[3] Olmos Pineda, I., González-Bernal, J. A., Minería de Datos, Universidad Politécnica de Puebla, México (2007).

[4] Martínez, M. D., *Minería de datos*, Universidad Nacional del Noroeste Facultad de Ciencias Exactas, Naturales y Agrimensura, Argentina, (2006).

[5] Reyes Saldaña, J. F., García Flores, R. *El proceso de descubrimiento de conocimiento de bases de datos.* Revista Ingenierías **VIII**, No. 26, pp 37-47 (2005).

[6] Ballesteros Román, A., *Minería de Datos Educativa Aplicada a la Investigación de Patrones de Aprendizaje en Estudiante en Ciencias*, Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional, México (2012).

[7] Gómez Arenas, L. I., *Evaluación Comparativa de Herramientas para la Minería de Datos y sus Aplicaciones*, Instituto Tecnológico de León, Guanajuato, México (2005).

[8] Cristobal, R., Sebastian, V., Mykola, P. & Baker R., *Handbook of Educational Data Mining*, CRC Press; 1st. Edition, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, (2010).

[9] Han, J., Kamber, M., Data *Mining: Concepts and Techniques 2nd. Edition.* Morgan Kaufmann Publishers; (The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, USA, 2006).

[10] Romero Morales,C., Ventura Soto, S., Hérvas Martínez, C. *Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web*, Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, (2005), pp.49-56.

[11] Luan J., *Aplicaciones de Minería de datos en la Educación Superior*, (IBM Press and IBM Corporation, Estados Unidos de America, 2010).

[12] Peña-Ayala, A., *Educational data mining: A survey and a data mining-based analysis of recent Works*, WOLNM & ESIME Zacatenco, Instituto Politécnico Nacional, México (2013).