***22-11-2021***

**Miguel Ernesto Campos Ramos**

***Tipos de modelos Minería de datos***

***Compresión del negocio***

Compresión de negocio es la fase con la que se abre proceso. Se encuentra enfocada en la compresión de los objetivos y exigencias de proyecto partiendo desde la perspectiva del negocio.

Tras ello, es necesario adquirir este conocimiento de los datos repetimos, siempre desde el punto de vista del negocio y convertirlo de la definición de un problema de minería de datos, trazando un plan preliminar de acuerdo a los objetivos planificados.

***Comprensión de los datos***

Compresión de los datos comienza con la colección de datos inicial para continuar con las actividades que permiten alcanzar una familiaridad con ellos que permita identificar los problemas de calidad de datos.

La calidad de datos tiene varias dimensiones exactitud que se refleje lo que esta

pasando, totalidad que se encuentren los datos completos en el sistema, oportunidad accesible cuando sea necesario, relevancia, nivel de detalle y consistencia mismo datos en todas las áreas o sistemas, por lo que será necesario comprobar como son los datos en cada una de estas dimensiones.

De esta manera ya se puede empezar a descubrir los primeros conocimientos en los datos, detectar aquellos conjuntos o subconjuntos de información interesantes para formar hipótesis que permitan avanzar en el descubrimiento de información oculta.

***Preparación de datos***

Preparación de datos se quieren cubrir todas las actividades necesarias para adaptar los datos origen en bruto y aproximarlos al conjunto de datos final los datos que serán fuente de las herramientas de modelado.

Las tareas de preparación o de limpieza de datos van a ser realizadas repetidas veces y no en cualquier orden. Entre estas tareas tenemos la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para su preparación para las herramientas de modelado.

***Modelado***

Como veremos en el próximo apartado, existen múltiples técnicas de modelado de datos, siendo en esta fase del proceso cuando, tras el conocimiento adquirido, se seleccionan las adecuadas siempre de acuerdo a los objetivos de negocio y del proyecto y se aplican

**Clases**: Las observaciones se asignan a grupos predeterminados.

**Clusters**: Se construyen grupos de observaciones similares según un criterio prefijado.

**Asociaciones**: Las observaciones son usadas para identificar asociaciones entre variables.

**Patrones secuenciales**: Se trata de identificar patrones de comportamiento y tendencias.

La caracterización o resumen, la discriminación o contraste, el análisis de asociación, las clasificaciones, la predicción, el clustering o la detección de agrupamientos, la detección de anomalías, el análisis de la evolución y de la desviación.

También existen diferentes técnicas para resolver el mismo tipo problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos para adecuar estos a la técnica pues ser necesario.

***Evaluación***

En estas etapas en el proyecto ya se ha construido un modelo. Para asegurarnos de que se cumple con los estándares de calidad propuestos para el proyecto es necesario evaluarlo desde una perspectiva de análisis de datos. Es decir, antes del proceder al despliegue final y su puesta en producción, es importante realizar una batería de pruebas junto con la revisión de cada paso ejecutados en la creación del modelo, que ayude a comparar el modelo obtenido con los objetivos de negocio.

Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en minería de datos.

Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de minería de datos.

Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

Análisis de los resultados. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Las redes neuronales

Se trata de un paradigma de aprendizaje y de un procesamiento automatizado que se encuentra inspirado en la manera que funciona el sistema nervioso de los animales.

Es un sistema que permite interconectar las neuronas en una red (red neuronal) que presta colaboración para la producción de estímulos de salida.

Entre los ejemplos de las redes neuronales se pueden mencionar:

La perceptrón (tipo de red neuronal artificial)

La perceptrón multicapas

Los mapas autoorganizados

La regresión lineal

Es una de las técnicas más utilizadas para la formación de relaciones entre datos. Se trata de un sistema rápido y eficaz pero que a su vez cuenta con insuficiencias en los espacios multidisciplinarios donde puedan relacionarse más de 2 variables.

Los árboles de decisión

Se trata de un modelo de predicción que se utiliza en el campo de la inteligencia artificial a partir de una base de datos en donde se construyen los diagramas de construcciones lógicas. Es un sistema similar a la predicción que está basado en reglas que sirven para la representación de una serie de condiciones que se presentan de manera sucesiva en la resolución de problemas.

Los modelos estadísticos

Se trata de una expresión simbólica en forma de igualdad que es empleada en los diseños experimentales y en la regresión, con la intención de identificar los factores que modifican la variable de respuesta

El agrupamiento

Consiste en la agrupación de una serie de vectores según determinados criterios que habitualmente son a distancia. Se trata de la disposición de los vectores de entrada de manera que estén más cercanos a los que tengan características comunes.

Según el objetivo que tenga la realización del análisis los algoritmos se pueden clasificar como algoritmos supervisados, que predicen un dato desconocido inicialmente a partir de otros datos que son de conocimiento previo. Y los algoritmos no supervisados, los cuales descubren patrones y tendencias que se presentan los datos.