Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de datos

Docente: Mayra Cristina Berrones Reyes

Resúmenes de técnicas de minería de datos

Grupo 002

Sofía Pamela Rosales Garza 1799219

San Nicolás de los Garza, 2 de octubre del 2020

Podemos dividir las técnicas de minería de datos en dos categorías, descriptivo y predictivo. El objetivo de las técnicas descriptivas es encontrar patrones que den un resumen de las relaciones ocultas dentro de un conjunto de datos y nos permiten descubrir sus características más importantes. Por otro lado, las técnicas predictivas, como su nombre lo dice, predicen el valor de un atributo en particular basándose en los datos recolectados de otros atributos, es decir, existe una variable dependiente y una o muchas independientes.

1. Reglas de asociación

La primera técnica de minería de datos que vimos en las exposiciones de nuestros compañeros fue la de reglas de asociación. Esta técnica pertenece a la categoría descriptiva. Permite descubrir hechos que ocurren en común dentro de un conjunto de datos, se podría decir que extrae información por coincidencias entre los datos cumpliendo con su objetivo de encontrar relaciones entre ellos. Entre sus aplicaciones tenemos definir patrones de navegación dentro de una tienda, promocionar pares de productos en un supermercado, ayudar a la toma de decisiones, analizar la información de ventas, distribuir mercancía de cierta forma para ayudar a que las ventas incrementen, segmentar clientes basándonos en sus patrones de compra.

Existen diversos tipos de reglas de asociación, entre ellos la asociación cuantitativa, multidimensional y de multinivel; que a su vez se dividen en Booleana (presencia o ausencia de un ítem) y cuantitativa (describe asociaciones entre ítems cuantitativos), unidimensional y multidimensional, y de un nivel y multinivel, respectivamente.

Dentro de las reglas de asociación existen tres métricas de interés: soporte, confianza y lift. El soporte de una regla “Si A => B” se define como el número de veces o la frecuencia (relativa) que A y B aparecen juntos en una base de datos de transacciones, o sea, que sería la intersección entre A y B dentro del conjunto. Dada la misma regla, su confianza es igual al cociente del soporte de la regla y el soporte del antecedente solamente [Soporte (A => B)] / [Soporte (A)]. Finalmente, el lift refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente; es el cociente de la intersección de A y B, y la probabilidad de A multiplicada por la probabilidad de B.

2. Detección de outliers

La detección de outliers, o detección de datos atípicos, es una técnica de minería de datos descriptiva, estudia a las observaciones que se desvían mucho del resto de los elementos pareciendo sospechosas, este tipo de observaciones pudieron ser generadas por mecanismos diferentes al resto de los datos del conjunto; en otras palabras, la detección de outliers estudia el comportamiento de valores que difieren del patrón general de una muestra.

La técnica de detección de outliers puede ser aplicada en sectores financieros, de seguridad, etc. Entre sus aplicaciones tenemos el aseguramiento de ingresos en las telecomunicaciones, la detección de fraudes financieros y la seguridad y detección de fallas de cualquier tipo en un sistema.

Esta técnica se lleva a cabo realizando pruebas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos.

Después de confirmar que los datos atípicos no se deben a un error al momento de construir la base de datos, eliminarlos no es la solución. Eliminarlos o sustituirlos puede modificar las inferencias que se realicen a partir de esa información, debido a que introduce un sesgo. La mejor solución sería quitarle peso a los outliers mediante técnicas robustas, que son técnicas que se ven medos afectadas por variaciones respecto a las hipótesis de los modelos.

3. Regresión

Regresión es la técnica de minería que mi equipo y yo escogimos para exponer en clase. Recientemente vimos este tema en materias como métodos estadísticos y estadística aplicada. En lo personal, el maestro que me dio la clase de métodos estadísticos nos enseñó un poco de Jupyter, y dentro de esa enseñanza estaba cómo hacer regresiones en esta plataforma, por lo que hacer el problema del ejemplo de regresión lineal simple que expusimos fue más sencillo teniendo esas bases y también para el problema de regresión de la tarea de los ejercicios 1. También en estadística aplicada hemos visto regresiones lineales pero en lugar de Jupyter usamos R.

La regresión es una técnica de minería de datos de la categoría predictiva, es decir, predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

La regresión se encarga de realizar un análisis del vínculo entre una variable dependiente “y” y una o varias variables independientes “x”, para de esta manera, encontrar una relación matemática. El que existan una o más variables independientes da pie a que podamos diferenciar el modelo de regresión lineal simple y el modelo de regresión lineal múltiple.

Hablando del modelo de regresión lineal simple, como ya mencioné, se trata de una sola variable regresora y su modelo está definido por $y = \beta_o + \beta_1 x + e$.

Por otro lado, tenemos el modelo de regresión lineal múltiple, se relaciona la variable dependiente “y” con los k regresores o variables predictivas, el modelo que lo describe es $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$.

La regresión puede ser aplicada en distintos y variados ámbitos, entre ellos la medicina, informática, estadística, comportamiento humano, industria, entre otros.

4. Predicción

La predicción es una técnica de minería de datos que como su nombre lo dice, es del tipo predictiva. Esta técnica analiza datos actuales históricos reales para obtener información sobre acontecimientos no conocidos o futuros. Antes de cualquier cosa, tenemos que trabajar con un buen modelo de predicción y para esto se necesita definir adecuadamente el problema que se tenga, mencionando el objetivo, las salidas deseadas, las condiciones que tiene, etc., evidentemente tenemos que recopilar datos para nuestra base, elegir una medida o indicador de éxito y preparar los datos, o sea, tratar con campos vacíos, con valores categóricos, entre otros.

Después de esto es importante dividir nuestros datos, el 70% se destina a un conjunto de entrenamiento, el 15% a un conjunto de validación y por último, el 15% restante se reserva para el conjunto de pruebas.

En esta técnica se usan árboles de diferentes tipos, el primero que mencionaré es el árbol de decisión. Trata de un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.

Para dividir el espacio muestral es preciso aplicar una serie de reglas o decisiones para que cada región contenga una buena proporción. Si una subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos de la misma clase.

Los árboles de decisión se pueden clasificar en dos tipos:

- Árboles de regresión, la variable respuesta “y” es cuantitativa
- Árboles de clasificación, la variable respuesta “y” es cualitativa

La estructura correcta de un árbol de decisión está formada por nodos y se leen de arriba hacia abajo. Hasta arriba tenemos el primer nodo o nodo raíz, siendo la variable más importante; después los nodos internos o intermedios, que vuelven a dividir el conjunto de datos en función de las variables; y finalmente los nodos terminales u hojas, que se ubican en la parte inferior de la estructura del esquema y cumplen con la función de indicar la clasificación definitiva.

Como mencioné antes, los árboles de decisión se subdividen en árboles de clasificación y de regresión. Los árboles de clasificación hacen preguntas del tipo ¿x_k <= c? para las covariables cuantitativas o preguntas del tipo ¿c_k = $nivel_j$? para las covariables cualitativas. El árbol de regresión consiste en hacer preguntas de tipo ¿x_k <= c? para cada una de las covariables, de esta forma el espacio de las covariables es dividido en hiper-rectángulos y todas las observaciones que quedan dentro de uno tendrán el mismo valor estimado $\hat{y}$.

Hay dos tipos de nodo, los nodos decisión que tienen una condición al principio y tienen más nodos debajo de ellos, y los nodos de predicción o nodos hijo que no tienen ninguna condición ni nodos debajo de ellos. La información de cada nodo es su condición (si se toma una decisión o no), gini (es una medida de impureza), samples (número de muestras que satisfacen las condiciones para llegar a ese nodo), value (cuántas muestras llegan a ese nodo) y class (qué clase se les asigna a las muestras que llegan a ese nodo).

5. Clustering

El clustering es una técnica de la categoría descriptiva. Es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear grupos basándonos en sus características similares.

Dentro de las aplicaciones del clustering tenemos investigación de mercado, identificar comunidades, prevención del crimen, procesamiento de imágenes, transformación de datos como variables cuantitativas, variables binarias, variables categóricas, entre muchas otras.

Existen cuatro tipos básicos de análisis, el primero es el Centroid Based Clustering, donde cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más usado de este tipo es el de k-medias.

Después tenemos el Connectivity Based Clustering, los clusters se definen agrupando a los datos más similares o cercanos (los puntos más cercanos están más relacionados que otros puntos más lejanos). La característica principal es que un clister contiene a otros clusters (representan una jerarquía). Un algoritmo usado de este tipo es Hierarchical clustering.

Luego está el Distribution Based Clustering. En este método cada cluster pertenece a una distribución normal, la idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. Un algoritmo de clustering perteneciente a este tipo es Gaussian mixture models.

Finalmente, está el Density Based Clustering. Los clusters son definidos por áreas de concentración. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas ente clusters.

6. Visualización

La visualización es una técnica descriptiva, es la representación gráfica de la información y los datos más importantes en una base de datos.

Al utilizar elementos visuales, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. Esta técnica es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Existen muchas técnicas y aproximaciones para la visualización según la naturaleza o clasificación del dato de la información. Según la complejidad y elaboración de la información podemos tener la siguiente clasificación: elementos básicos de representación de datos, cuadros de mando e infografías.

Los elementos básicos de representación de datos son el caso más sencillo, dentro de las visualizaciones básicas tenemos gráficas: barras, líneas, columnas, puntos, pastel; mapas: burbujas, coropletas (mapa temático), de calor, de agregación; y tablas: con anidación, dinámicas, de drill-down, de transiciones, entre otras. Un cuadro de mando es una composición completa de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas.

Las infografías no están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos, es decir, las infografías se utilizan para "contar historias" que son contadas mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como símbolos, leyendas, dibujos, imágenes, etc.

7. Patrones secuenciales

Los patrones secuenciales es una técnica predictiva. Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias. Es una clase especial de dependencia en las que el orden de acontecimientos es considerado. El patrón secuencial describe el modelo de compras que hace un cliente particularmente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo. Son eventos que se enlazan con el paso del tiempo.

Esta técnica trata de buscar asociaciones de la forma "si sucede x evento en el tiempo t entonces sucederá el evento y en el instante t+n". Su objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos. Utiliza reglas de asociación secuenciales, reglas que expresan patrones de comportamiento secuencial, que se dan en distintos instantes en el tiempo.

Las características principales de los patrones secuenciales son:

- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes (patrones secuenciales) son las subsecuencias de una secuencia que tienen un soporte mínimo

Podemos encontrar aplicaciones de esta técnica en áreas como medicina, biología, bioingeniería, en la web, análisis de mercado, distribución y comercio, aplicaciones financieras y banca, aplicaciones de seguro y salud privada y en los deportes. Se desarrolla en tipos de base de datos temporales, documentales y relacionales.

8. Clasificación

Dentro de la categoría predictiva de las técnicas de minería de datos se encuentra la clasificación. Esta técnica es una de las que usamos más comúnmente, organiza o mapea un conjunto de atributos por clase dependiendo de sus características.

Su funcionamiento consiste en estimar un modelo usando los datos recolectados en nuestra base para poder hacer predicciones futuras.

Dentro de la clasificación existen varios tipos, entre ellos el Support Vector Machines (SVM), la clasificación basada en asociaciones, la regla de Bayes, que es un tema que ya conocemos desde que llevamos la materia de probabilidad básica; las clasificaciones neuronales, consisten en tres capas: de entrada, oculta y salida, trabajan directamente con números y en caso de que se desee trabajar con datos nominales deben ser numerados. Además de usarse en clasificación también se usan en regresión y agrupamiento.

También existen los árboles de decisión, que ya definimos lo que son en el tema de predicción, pero repito que son una serie de condiciones organizadas que su estructura tiene un orden jerárquico, como su nombre lo dice, en forma de árbol. Además de ser mencionados en predicción, también son utilizados en las técnicas de agrupamiento y regresión como las clasificaciones neuronales.