UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Resúmenes

Técnicas de Minería de Datos

MIRTHALA NALLELY CANTÚ CORTINA
1614768

02-OCTUBRE-2020

Técnicas de Minería de Datos: OUTLIERS

En esta técnica se nos mencionó que la podemos utilizar para detectar algunos datos raros o atípicos, diferentes al patrón que sigue el resto; de esta forma podemos analizar estos datos y descubrir qué pasó en esas situaciones específicas.

Aquí podremos analizar alguna observación que se desvía mucho del resto de las observaciones, ya que esta pudo haber sido generada por mecanismos diferentes al resto y de esta forma no se estaría considerando correctamente al conjunto de datos.

Esta técnica se puede utilizar en diferentes situaciones, entre estas, puede ser en alguna empresa para detectar fraudes financieros, en alguna tienda para revisar inventarios o ingresos y egresos, etc.

Todo esto se lleva a cabo mediante pruebas estadísticas no paramétricas, para la comparación de resultados dependiendo de la capacidad de detección de los algoritmos utilizados, es una herramienta que muchas empresas o negocios ponen en práctica para evitar las pérdidas, por lo que es importante tenerla en cuenta y aprender sobre esto.

Técnicas de Minería de Datos: Predicción

Un árbol de decisión es un modelo predictivo que nos agrupará observaciones que tengan valores similares para la variable dependiente, de esta forma se mencionó que se dividirá el espacio muestral en otras subregiones para poder estudiarlo y analizarlo de mejor manera. Estos árboles pueden ser de regresión; es decir, con respuesta cuantitativa, o de clasificación, con respuesta cualitativa.

El árbol se divide a través de nodos, y estos pueden ser de decisión (con una condición al principio y más nodos debajo de ellos) o de predicción (sin condición ni nodos debajo). La información que corresponde a cada nodo del árbol puede ser de condición (si es un nodo donde se toma alguna decisión), gini (esto es una medida de impureza), samples (es el núnero de muestras que satisfacen las condiciones que fueron necesarias para llegar hasta el presente nodo), Value (indica cuántas muestras de cada clase llegaron hasta el presente nodo) y class (que nos indica qué clase se les asigna a las muestras que llegan hasta este nodo). Cuando gini vale 0 significa que este nodo es totalmente puro.

El árbol de regresión es para dividir el espacio y que todas las observaciones que queden dentro de un hiper-rectángulo tengan el mismo valor estimado, es fácil de entender o interpretar, no se necesita mucha preparación de los datos, se pueden manejar covariables tanto cualitativas como cuantitativas, y no exige supuestos distribucionales; por lo que resulta ventajoso utilizarlo.

Los bosques aleatorios es una técnica basada en árboles de decisión, éste obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar.

Técnicas de Minería de Datos: Reglas de asociación.

Estas reglas se derivan de un tipo de análisis que extrae información por medio de coincidencias, y así poder encontrar relaciones dentro de un conjunto de movimientos o transacciones en caso de que las haya.

Utilizar las reglas de asociación nos va a permitir encontrar combinaciones de artículos que con mayor frecuencia ocurren en una base de datos, así podremos por ejemplo en un súper mercado, saber qué es lo que la gente acostumbra llevar, qué productos se buscan junto a otros con frecuencia, y anticipar lo que el cliente llevará, poniéndolo a su alcance.

Las aplicaciones son para predecir patrones dentro de una tienda, para colocar promociones en base a ciertos productos, analizar la información de las ventas, distribuir bien la mercancía en las tiendas y segmentar los clientes en base a sus patrones de compra.

Tenemos diferentes reglas de asociación: La cuantitativa, que describe asociaciones entre ítems cuantitativos o atributos, la asociación booleana hace asociaciones entre la presencia o ausencia de un ítem. La asociación multidimensional se realiza con base en las dimensiones de los datos que involucran regla de asociación unidimensional (si los ítems o atributos se referencian en una sola dimensión), o multidimensional (si se referencian en dos o más dimensiones).

La asociación multinivel involucra reglas de un nivel o multinivel (ítems referenciados a varios niveles de abstracción).

El soporte se define como el número de veces que los ítems A y B aparecen juntos en una base de datos. Se considera regla de bajo soporte cuando se dice que aparecieron por casualidad.

El lift refleja el aumento de la probabilidad de que ocurra la compra B, dado que sabemos que ocurrió A.

Técnicas de Minería de Datos: Clustering

El clustering es una técnica de máquina no supervisada, esto consistirá en agrupar puntos de datos para así crear participaciones basadas en similitudes. Tiene distintas áreas de uso, se puede utilizar para una investigación de mercado, para identificar comunidades determinadas, para la prevención de un crimen, para el procesamiento de imágenes, etc.

En el centroid based clustering cada cluster es representado por un centro, éstos clusters se construyen basados en la distancia de punto de cada dato hasta el centro.

En el connectivity based clustering los clusters se definen agrupando datos con mayor similitud (que son los más cercanos). Aquí se representan jerarquías, ya que un cluster contiene a otros clusters.

En la distribution based clustering cada cluster pertenece estrictamente a una distribución normal, aquí los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución.

En el density based clustering los clusters son definidos por áreas de concentración, se conectan puntos con poca distancia entre ellos.

El método K-Medias es un algoritmo muy utilizado, basado en centroides, donde k nos indicará el número de clusters e lo definirá el usuario. A medida que K aumenta, la varianza de cada cluster disminuye.

El método del codo consta de graficar la reducción de la varianza total a medida que K aumenta.

Técnicas de Minería de Datos: Visualización de datos.

En este tema se nos habló sobre la representación gráfica de la información y datos. Para esto se pueden utilizar elementos visuales como mapas, gráficas o cuadros; gracias a esto se pueden comprender y detectar mejor las tendencias, valores atípicos o patrones que se formen en un conjunto o base de datos.

Existen varios tipos de visualizaciones:

Los elementos básicos de representación de datos pueden ser gráficas (de barras, de pastel, columnas o puntos, etc.), mapas y tablas.

Los cuadros de mando son una composición de visualizaciones individuales pero que se relacionan entre ellas, se utilizan más que todo para el análisis de conjuntos de variables y toma de decisiones.

Las Infografías no tienen como objetivo el análisis de las variables sino a la construcción de narrativas a partir de los datos, esta narrativa se construye a través de la disposición de la información combinada con símbolos, leyendas, dibujos, etc.

En cualquier empleo es de suma importancia el visualizar los datos de manera adecuada, de tal manera poder tomar decisiones lo más acertadas posible y usar elementos visuales para contar historias con los datos para informar; la visualización de datos se encuentra justo en el centro del análisis y la narración visual.

Técnicas de Minería de Datos: REGRESION

La primera forma de regresión lineal que se tuvo fue el método de los mínimos cuadrados, y fue publicado por Legendre. Después Gauss desarrolló un poco más profundamente el método, y a parte incluyó una versión del teorema de Gauss-Markov.

Estos modelos lineales son una aplicación muy ágil y simple de la realidad que se da por parte de la matemática y la estadística para facilitarnos el trabajo.

Esta técnica es de la categoría predictiva, predice el valor de un atributo en particular, basándose en los datos recolectados de otros atributos. Analiza el vínculo entre una variable que es dependiente y una o más variables independientes, de esta forma crea una relación matemática.

Vimos dos tipos de regresión, la regresión lineal simple y la regresión múltiple.

La regresión lineal simple se da cuando el análisis de regresión se trata solo de una variable regresora. Su forma es la siguiente: $y = \beta_o + \beta_1 x + e$, donde e es una variable aleatoria.

En la regresión lineal múltiple se habló sobre un modelo que se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos.

Estos modelos se pueden utilizar en distintas y variadas áreas como lo son la medicina, la informática, estadística, para analizar el comportamiento humano, en la industria, etc.

Técnicas de Minería de Datos: CLASIFICACION

Es la técnica que más comúnmente se utiliza, ésta organiza o mapea un conjunto de datos o atributos por clases, dependiendo de sus características en común. Aquí se estima un modelo usando los datos recolectados para hacer predicciones a futuro.

Este método se puede utilizar en diversas áreas, ya que el clasificar datos con características similares por lo regular o en la mayoría de los casos es muy útil, por dicha razón esto se aplica tanto en lo profesional como en lo personal.

Existen distintas técnicas de clasificación, entre ellas se encuentran la clasificación por inducción de árbol de decisión, la clasificación bayesiana, las redes neuronales, Support Vector Machines y la clasificación basada en asociaciones.

Las redes neuronales trabajan directamente con números y si se desea trabajar con datos nominales pues éstos deben enumerarse. No solo se usan en la clasificación, sino que éstas se pueden aplicar también en agrupamiento, la regresión, etc.

Las redes neuronales consisten o constan de tres capas: de entrada, oculta y de salida.

El árbol de decisión se da por una serie de condiciones organizadas de forma jerárquica, son muy útiles cuando se mezclan datos categóricos o numéricos.

Técnicas de Minería de Datos: PATRONES SECUENCIALES

Estos patrones secuenciales se especializan principalmente en analizar datos y encontrar sub-secuencias interesantes dentro de un grupo de secuencias.

Es una clase de dependencia en las que el orden de acontecimientos es considerado. Describe el modelo de compras que hace un cliente en particular o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo. Todos estos eventos se enlazan con el paso del tiempo.

En este método se trata de buscar asociaciones o similitudes “si sucede x en el tiempo t, entonces sucederá y un tiempo determinado después”.

La finalidad es el poder describir concisamente las relaciones temporales que hay entre los valores de los atributos de un conjunto de ejemplos.

Se usan reglas de asociación secuencial, reglas que nos expresan patrones de comportamiento secuencial, se dan en instantes distintos de tiempo.

En ésta técnica, el orden importa, se tiene la finalidad de encontrar patrones. Una secuencia es una lista en orden de ítems.

Se tienen diferentes áreas de aplicación, como lo son la medicina, la biología, el análisis de mercado, finanzas y banca, seguro y salud privada, deportes, bases de datos, etc.

El agrupamiento de patrones secuenciales es la tarea de separar en grupos los datos, de forma que se junten los que tienen características similares entre sí.

La clasificación con datos secuenciales se da cuando los datos contiguos tienen alguna relación, se expresan patrones de comportamiento secuenciales.