UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Minería de Datos

RESUMENES

Omar Alejandro Garza Espinosa 1931548

Profesor: Mayra Cristina Berrones Reyes
Grupo: #03

Viernes 2 de octubre del 2020

Ciudad Universitaria, San Nicolás de los Garza, N.L.

Resúmenes de las Técnicas de Minería de Datos

Reglas de Asociación

Esta técnica de minería de datos busca patrones, asociaciones, correlaciones o estructuras similares entre conjuntos de elementos u objetos en bases de datos de transacciones, bases de datos relacionales y otros repositorios de información disponibles.

Aplicaciones:

- Análisis de Datos de la Banca: brinda un mejor servicio hacia los clientes según a los productos financieros que han utilizado o las herramientas que usan en la banca.
- Cross-Marketing: se utiliza mucho para mejorar las ventas en una empresa ya que relacionan productos que les puede interesar comprar a los clientes.
- Diseño de Catálogos: se utiliza para que al momento de realizar un catálogo, éste sea atractivo para el consumidor.

Un ejemplo sencillo puede ser cuando las personas compran pan en una tiendita, por lo general también compran leche. Otro seria si una persona compra soda, por lo general compra también frituras.

Conceptos importantes en las reglas de asociación:

1. Soporte: Fracción de transacciones que contiene un itemset.
2. Conjunto de elementos frecuente: Un conjunto de elementos cuyo soporte es mayor o igual que un umbral de mínimo.
3. Conjunto de elementos: Una colección de uno o más artículos.
4. Recuento de Soporte: Frecuencia de ocurrencia de un itemset.
5. Confianza: Mide que tan frecuente ítems en Y aparecen en transacciones que contienen X.

Objetivo:

El objetivo de la minería de reglas de asociación es encontrar todas las reglas o patrones de una base de datos teniendo en cuenta:

- Umbral mínimo de soporte
- Umbral mínimo de confianza

Tipos de Reglas de Asociación:

- Enfoque de 2 pasos : se dice a dos pasos porque primero se generan los elementos frecuentes y después se hacen las reglas de asignación.
- Principio “a priori”: si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes.

Clasificación

Es una técnica de la minería de datos utilizada para el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene una base de datos.

La clasificacion se encuentra dentro de las tareas predictivas ya que trata de acertar a lo que va a pasar en un futuro de un atributo en particular basándose en los datos recolectados de otros atributos.

Los métodos más comunes son:

- Análisis discriminente: es utilizado para encontrar combinaciones lineales y así separar los datos.
- Reglas de clasificación: busca coincidencias en los datos y los clasifica en uno mismo.
- Arboles de decisión: se hace un esquema que ayude a la toma de decisiones y así clasificar los datos.
- Redes neuronales artificiales: es utilizada con unidades conectadas para transmitir señales y así clasificar datos.

Características:

- Precisión en la predicción
- Eficiencia
- Robustez
- Escalabilidad
- Interpretabilidad

Un ejemplo sencillo puede ser clasificar a un grupo de alguna materia según la calificacion obtenida en algún parcial ya presentado.

Outliers

La técnica de Detección de Outliers es una técnica de minería de datos que estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

Valor atípico:

Son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por: errores de entrada de datos, acontecimientos extraordinarios, valores extremos y causas no conocidas. Los datos atípicos distorsionan los resultados de los análisis por lo que hay que identificarlas y tratarlos de manera adecuada, sino se puede caer en un mal análisis o una mala predicción.

Los tipos de técnicas para identificar datos atípicos son:

- Métodos univariantes de detección de outliers
- Métodos multivariantes de detección de outliers

Técnicas para la detección de valores atípicos:

- Prueba de Grubbs
- Prueba de Dixon
- Prueba de Tuckey
- Análisis de Valores y Atípicos de Mahalanobis
- Regresión Simple

Una vez detectados los valores atípicos, se pueden eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variables. Lo mejor sería quitarle peso a esas observaciones atípicas mediante técnicas rubustas.

Aplicaciones:

- Detección de fraudes financieros por si se sospecha de algún movimiento o cantidad extraña en cuentas de clientes.
- Tecnología Informática al momento de hacer estudios de mercado o algún proyecto específico.
- Nutrición y salud al encontrar datos que estan afectando o simplemente no causan alguna alteración a una enfermedad encontrada.
- Negocios al observar datos que estan afectando en las finanzas de la empresa o a la venta de algún producto, incluso se puede utilizar en producción al encontrar insuficiencias.

Patrones Secuenciales

Conceptos:

- Minería de datos secuenciales: Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo ya que como vayan ocurriendo se van conectando.
- Reglas de asociación secuencial: Expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

Características de los patrones secuenciales:

- El orden de aparición importa.
- Su objetivo es encontrar patrones secuenciales.
- El tamaño de una secuencia es la cantidad de elementos en la base de datos.
- La longitud de la secuencia es la cantidad de ítems.
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.

- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Ventajas de los patrones secuenciales:

- Flexibilidad
- Eficiencia

Desventajas de los patrones secuenciales:

- Utilización
- Sesgado por los primeros patrones

Aplicaciones:

- Medicina: Predecir si un compuesto químico causa cáncer o no.
- Análisis de Mercado: Comportamiento de compras de los consumidores en tiendas comerciales.
- Web: Reconocimiento de spam de un correo electrónico.

Predicción

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento. En muchos casos, mediante los datos históricos de cierta cosa es suficiente para reconocer y comprender su vida o trayectoria y asi poder hacer una predicción precisa de lo que sucederá en el futuro. Existen dos tipos de variables: variables independientes que nos indica los datos ya conocidos y las variables de respuesta que son las variables a las que queremos llegar y en base a esto hacer una predicción.

Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para el uso de la predicción. Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos.

Aplicaciones:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro
- Predecir si va a llover mediante a como a llovido en años anteriores.
- Predecir el precio de venta de una propiedad en base a estudios de este mercado.
- Predecir en cualquier evento deportivo mediante resultados pasados.

Técnicas:

- Modelos estadísticos – regresión simple
- Estadística no lineal
- Redes Neuronales

Todo esta basado en ajustar una curva a través de los datos, es decir, encontrar una relación entre los predictores y los pronosticados y asi poder hacer una predicción.

Tipos de métodos de regresión:

- Regresión lineal
- Regresión lineal multivariante
- Regresión no lineal
- Regresión no lineal multivariante

Regresión

Una Regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir conocer si existe relación entre ellas. Con esto, podemos ajustar las variables a un modelo y asi poder hacer una predicción de lo que puede pasar en un futuro.

Existen dos tipos de regresión:

- Regresión lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.
- Regresión lineal multiple: cuando dos o más variables independientes influyen sobre una variable dependiente.

En Minería de Datos la Regresión se encuentra dentro de la categoría Predictivo.
Esta categoría tiene como objetivo analizar los datos de un conjunto y en base a eso, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

Análisis de regresión:

Permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés. El análisis de regresión nos permite explicar un fenómeno y predecir cosas acerca del futuro, por lo que nos será de ayuda para tomar decisiones y obtener los mejores resultados.

Existen dos tipos de variables:

- Variable dependiente: es el factor más importante, el cual se está tratando de entender o predecir.
- Variable independiente: es el factor que tú crees que puede impactar en tu variable dependiente.

Visualización de datos

Sirve para representar gráficamente los elementos mas importantes de una base de datos ya que utilizan elementos visuales como cuadros, gráficos o mapas los cuales proporciona una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos.

Tipos de visualización de datos:

- Gráficos: es el más común y es para presentar la información de manera sencilla como Gráficos Circulares, Líneas, Columnas y Barras aisladas o agrupadas, Burbujas, áreas, Diagramas de Dispersión y Mapas de tipo Árbol.
- Mapas: es una manera fácil de analizar los datos según a un país o ver en que lugar se concentra más cierta característica.
- Infografias: colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente.
- Cuadros de mando: es una herramienta que permite saber en todo momento el estado de los indicadores del negocio.

Aplicaciones:

- Comprender la información con rapidez: se puede analizar una gran cantidad de información importante de manera sencilla y agradable.
- Identificar relaciones y patrones: en algunas ocaciones, al ver la información de muchos datos en una grafica puedes ver la relación que hay unos con los otros.
- Identifique tendencias emergentes: con las graficas se puede observar las tendencias que podra tener algun activo y asi hacer una proyección o tener alguna ventaja del mercado.

Clustering

Se trata del proceso de dividir los datos en grupos de objetos similares. Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información de las variables que pertenecen a cada objeto se mide la similitud o parecido entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases.

Conceptos:

- Cluster: es una colección de objetos de datos que son similares entre sí dentro del mismo grupo.
- Análisis de cluster: dado un conjunto de puntos de datos, se trata de entender su estructura encontrando similitudes entre los datos de acuerdo con las características recolectadas en los datos.

Aplicaciones:

- Estudios de terremotos
- Aseguradoras
- Marketing
- Planificación de la ciudad

Algoritmos de Clustering

- Simple K-Means: este algoritmo debe define el número de clusters que se desean obtener.
- X-Means: este algoritmo es una variante mejorada del K-Means. Su ventaja está en haber solucionado deficiencias presentadas en K-Means que es tener que seleccionar a priori el número de clusters que se deseen obtener, a X-Means se le define un límite inferior K-min (número mínimo de clusters) y un límite superior K-Max (número máximo de clusters) y este algoritmo es capaz de obtener en ese rango el número óptimo de clusters, dando de esta manera más flexibilidad.
- Cobweb: se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia.