UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO

MATEMÁTICAS

Minería de Datos

RESUMENES
GPO 003

ALUMNA: DANIELA GOVEA SERNA
MATRICULA: 1722714

FECHA DE ENTREGA: 02/10/2020

TÉCNICAS DESCRIPTIVAS

Clustering

También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de 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 que brindan las variables que pertenecen a cada objeto se mide la similitud 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.

Un **cluster** es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.
Análisis de cluster: dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

Aplicaciones:
Estudios de terremotos: los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.
Planificación de la ciudad: identificación de grupos de casas según su tipo de casa, valor, y ubicación geográfica.
Marketing: ayudar a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.
Aseguradoras: identificación de grupos de asegurados de seguros de automóviles con un alto costo promedio de reclamo.
Aseguradoras: identificación de grupos de asegurados de seguros de automóviles con un alto costo promedio de reclamo.

Algoritmos de Clustering:

Simple K-Means: Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar.

X-Means: Este algoritmo es una variante mejorada del K-Means. Su ventaja fundamental está en haber solucionado una de las mayores deficiencias presentadas en K-Means, el hecho de 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 al usuario.

Cobweb: Pertenece a la familia de algoritmos jerárquicos. Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia.

Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos.

EM: Este algoritmo pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos. Está clasificado como un método de particionado y recolocación, o sea, Clustering Probabilístico. Se trata de obtener la FDP (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos.

Reglas de Asociación

Búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales 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
-Cross-marketing
-Diseño de catálogos

Soporte: Fracción de transacciones que contiene un itemset.
Conjunto de elementos frecuentes: Un conjunto de elementos cuyo soporte es mayor o igual que un umbral de mínimo.
Conjunto de elementos: Una colección de uno o más. k-itemset, un conjunto de elementos que contiene k elementos.
Recuento de soporte: Frecuencia de ocurrencia de un itemsette.
Confianza (c): Mide que tan frecuente items en Y aparecen en transacciones que contienen X.

El Objetivo Dado un conjunto de transacciones T, el objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo:
• Umbral mínimo de soporte
• Umbral mínimo de confianza
Enfoque de fuerza bruta:
• Lista todas las reglas de asociación posibles.
• Compruebe el soporte y la confianza para cada regla.
• Elimine las reglas que fallan en los umbrales mínimos.

RAM: Enfoque 2 pasos
Generación de elementos frecuentes: Generar todos los conjuntos de elementos cuyo soporte ≥ min sup.
Generación de reglas: Generar reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuente.

Cada conjunto de elementos en la red es un conjunto de elementos frecuente candidato.
• Calcular el soporte de cada candidato escaneando la base de dato
• Empareja cada transacción con cada candidato.

Estrategias de generación de elementos frecuentes
• Reducir el número de candidatos (M)
• Reducir el número de transacciones (N)
• Reducir el número de comparaciones (NM)

Reduciendo el número de candidatos
Principio de Apriori: si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes

El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos. Esto se conoce como la propiedad anti-monótona de soporte

Algoritmo Apriori
El núcleo del algoritmo de Apriori.
• Utilizar los conjuntos frecuentes (k-1) para generar candidatos a k-items frecuentes.
• Utilizar el escaneo de la base de datos y la coincidencia de patrones para recoger los recuentos de los conjuntos de elementos candidatos.

Comprimir una gran base de datos en una estructura compacta de árbol de patrones frecuentes (FP-tree). Muy condensado, pero completo para la minería de patrones frecuentes. Evita costosos análisis de bases de datos. Utilice un método de minería de patrones frecuentes, basado en el árbol de FP. Una metodología de dividir y conquistar: descomponer las tareas de minería en los más pequeños.

Detección de Outliers

Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra

Un valor atipico 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 y procedimiento
-Acontecimientos extraordinarios
-Valores extremos y/o faltantes

-Causas no conocidas

Los datos atípicos distorsionan los resultados de los analisis y por esta razón hay que identificarlas y tratarlos de manera adecuada

Cálculo de valores atípicos

Existen distintos tipos de técnicas para detectarlos y se pueden dividir en dos categorías principales:
-Metodos univariantes de detección de outliers
-Metodos multivariantes de detección de outliers

Técnicas para la detección de valores atípicos
1.Prueba de Grubbs
2.Prueba de Dixon
3.Prueva de Tukey
4.Analsis de Valores
5.Regresion Simple

Se puede eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura en la medición de la variable.
Si no se debe a un error, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esta información debido a que:
-Introduce un sesgo
-Disminuye el tamaño muestral
-Puede afectar a la distribución y a las varianzas

La mejor opción es quitarles peso a esas observaciones atípicas mediante técnicas robustas

Aplicaciones de la minería de datos en outliers
-Detección de fraudes financieros
-Tecnología informática y telecomunicaciones
-Nutrición y salud
-Negocios

Visualización

Sabemos que el conjunto de datos es información que siempre ha existido, solo que con el tiempo hemos sido capaces de adquirir nuevas técnicas y/o herramientas que nos permiten no solo interpretar esos datos, sino que también darle un uso para nuestro beneficio. Una de estas técnicas es la visualización de datos, que nos sirve para representar gráficamente los elementos

más importantes de nuestra base de datos.
La visualización de datos es la presentación de información en formato **ilustrado o gráfico.** Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Tipos de visualización de datos
Es importante conocer que existen diferentes tipos de Visualización de datos ya que uno de los grandes retos que enfrentan los usuarios de empresas, es que tipo de elemento visual se debe utilizar para representar la información de la mejor forma. Aunque existen muchos tipos, mencionaremos los mas comunes:

Gráficos: Este es el tipo más común y conocido, que utilizamos en nuestro día a día con las hojas de cálculo, para representar datos de manera sencilla, como Gráficos Circulares, Líneas, Columnas y Barras aisladas o agrupadas, Burbujas, áreas, Diagramas de Dispersión y Mapas de tipo Árbol.
Mapas: Con la popularización de Google Maps y su conocida API (interfaz de programación para aplicaciones), todos conocemos la visualización de datos en mapas para conocer, por ejemplo, la localización de nuestra flota de vehículos en tiempo real o bien la de las tiendas de un supermercado o los cajeros automáticos de nuestro banco en un mapa.
Infografías: Una infografía es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente. son excelentes para ayudarnos a procesar más fácil, la información compleja.
Una infografía es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente. son excelentes para ayudarnos a procesar más fácil, la información compleja.
Cuadros de Mando (Dashboards): En el entorno empresarial, un cuadro de mando es una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos, etc. y que nos dice lo que está pasando en la empresa (idealmente en tiempo real) para poder tomar decisiones adecuadas, ya sean correctivas o de planeación.

La visualización de datos es la presentación de información en formato ilustrado o grafico. Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Aplicaciones
Comprender la información con rapidez: Mediante el uso de representaciones gráficas de información de negocios, las empresas pueden ver grandes cantidades de datos de formas claras y cohesivas y sacar conclusiones a partir de esa información.
Identificar relaciones y patrones: Incluso muy grandes cantidades de datos complicados comienzan a tener sentido cuando se presentan de manera gráfica; las empresas pueden

reconocer parámetros con una correlación muy estrecha.
Identifique tendencias emergentes: El uso de la visualización de datos para descubrir tendencias en los negocios y en el mercado puede dar a las empresas una ventaja sobre la competencia, y eventualmente tener un impacto en la base de operación.
Comunique la historia a otras personas: Una vez que una empresa ha descubierto nuevos insights a partir de la analítica visual, el paso siguiente consiste en comunicar esos insights a otras personas. En este paso es importante utilizar diagramas, gráficas u otras representaciones visualmente impactantes de los datos porque motiva la participación y transmite el mensaje con rapidez.

A medida que la "era del big data" entra en pleno apogeo, la visualización es una herramienta cada vez más importante para darle sentido a los billones de filas de datos que se generan cada día. La visualización de datos ayuda a contar historias seleccionando los datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos. Una buena visualización cuenta una historia, eliminando el ruido de los datos y resaltando la información útil

TÉCNICAS PREDICTIVAS

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.

Existen dos tipos de regresión:
1. Regresión Lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.
2. Regresión Lineal Múltiple: 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.

● Variable(s) dependiente(s): Es el factor más importante, el cual se está tratando de entender o predecir.
● Variable(s) independiente(s): Es el factor que tú crees que puede impactar en tu variable

dependiente.

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.

Clasificación

Es una técnica de la minería de datos donde el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene,.
La tarea es predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

Métodos:
Análisis discriminante: método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos
Reglas de clasificación: buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación
Arboles de decisión: método analítico que a través de una representación esquemática facilita la toma de decisiones

Redes neuronales artificiales (también conocido como sistema conexionista) es un modelo de unidades conectadas para trasmitir señales

Características de los métodos:
● Precisión en la predicción: Capacidad de predecir correctamente.
●Eficiencia: Costos computacionales.
● Robustez: Habilidad para funcionar con ruido y ausencia de ciertos valores.
● Escalabilidad: Habilidad para trabajar con grandes cantidades de datos.
● Interpretabilidad: Entendimiento y comprensión que brinda.

Patrones Secuenciales

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 El orden de acontecimientos es considerado.
Se busca asociaciones de la forma “si sucede de la forma X en el instante de tiempo t entonces sucederá en el evento Y en el instante t+n”. El objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

Reglas de asociación secuencial: Expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

Características

●El orden importa.
● Objetivo: encontrar patrones secuenciales.
● El tamaño de una secuencia es su cantidad de elementos.
● 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: Flexibilidad- Eficiencia
Desventajas: Utilización-Sesgado por los primeros patrones

Tipos de datos: ADN y Proteínas, Recorrido de clientes en un supermercado, Registros de accesos a una página web.

Aplicaciones:

Medicina: Predecir si un compuesto químico causa cáncer
Análisis de Mercado: Comportamiento de compras
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, el simple hecho de reconocer y comprender las tendencias históricas es suficiente para trazar una predicción algo precisa de lo que sucederá en el futuro.
Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo. Los valores son generalmente continuos. Las predicciones son a menudo (no siempre) sobre el futuro
Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos. Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos. Cuando este modelo se aplica a nuevas entradas de datos, el resultado es una predicción del comportamiento futuro de los mismos.

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 en función de la humedad actual
-Predecir el precio de venta de una propiedad
-Predecir la puntuación de cualquier equipo durante un partido de fútbol

Técnicas
La mayoría de las técnicas de predicción se basan en modelos matemáticos:
- Modelos estadísticos simples como regresión
- Estadísticas no lineales como series de potencias
- Redes neuronales, RBF, etc.

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

Tipos de métodos de regresión

Regresión Lineal
El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables.
En el Análisis de regresión simple, se pretende estudiar y explicar el comportamiento de una variable que notamos y, y que llamaremos variable dependiente o variable de interés, a partir de otra variable, que notamos x, y que llamamos variable explicativa, variable de predicción o variable independiente.

Regresión Lineal Multivariante
Permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta y, se determina a partir de un conjunto de variables independientes llamadas predictores x1, x2,x3,...
Es una extensión de la regresión lineal simple.
Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella

Regresión No Lineal univariable y multivariable
Método para encontrar un modelo no lineal para la relación entre la variable dependiente y un conjunto de variables independientes.
La regresión no lineal es una regresión en la que las variables dependientes o de criterio se modelan como una función no lineal de los parámetros del modelo y una o más variables independientes.
Se denomina regresión no lineal porque las relaciones entre los parámetros dependientes e independientes no son lineales

Redes neuronales

Utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión. Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.