UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Minería de Datos

Evidencia de Aprendizaje

Resumen de las técnicas de Minería de Datos

Docente: Mayra Cristina Berrones Reyes

Nombre: Gloria Nohemí Martínez Jiménez

Grupo:002 Matrícula: 1805800

01 de octubre de 2020

Clustering

El clustering es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándose en similitudes, es el análisis de grupos o agrupamiento es la tarea de agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo, llamado *clúster* sea más similar, en algún sentido u otro. Es la tarea principal de la minería de datos exploratoria y es una técnica común en el análisis de datos estadísticos. Los *clusters* se definen agrupando a los datos más similares o cercanos, los puntos más cercanos están más relacionados que otros puntos lejanos, la *característica principal* es que un cluster contiene a otros clusters, representan una jerarquía.

Cuando se representan la información obtenida a través de clusters se pierden algunos detalles de los datos, pero a la vez se simplifica dicha información.

Los clusters son definidos por *áreas de concentración*, se trata de conectar puntos cuya distancia entre si es considerada pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters. Para esta técnica solo disponemos de un conjunto de datos de entrada, sobre los que debemos obtener información sobre la estructura del dominio de salida, que es una información de la cual no se dispone.

Cada cluster está representado por un *centroide*, los clusters se construyen basados en la distancia de punto de los datos al centroide, se realizan varias iteraciones hasta llegar al mejor resultado, el algoritmo más usado es el de *k-medias*. Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Primeramente, se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones. Luego se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:

-Determina las coordenadas del centroide.

-Determina la distancia de cada objeto a los centroides.

-Agrupa los objetos basados en la menor distancia.

Finalmente quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

Reglas de asociación

Las reglas de asociación se definen como un conjunto de técnicas que permiten establecer relaciones de interés con la finalidad de descubrir hechos que aporten valor dentro de las variables que facilitan los datos que son enormes. Se derivan de un análisis que extrae información por coincidencias, su objetivo es encontrar relaciones dentro de un conjunto de transacciones. Estas reglas nos permiten encontrar las combinaciones que ocurren con mayor frecuencia en una base de datos y medir su fuerza e importancia.

Algunas aplicaciones de las reglas de asociación son la segmentación de clientes con base en patrones de compra, análisis de la información de ventas, distribución de mercancía en tiendas, promociones de pares de productos, entre otros.

Las reglas de asociación son: *asociación cuantitativa*, esta se basa en el tipo de valores que manejan, se divide en asociación booleana y asociación cuantitativa; *asociación multidimensional*, se basa en las dimensiones de los datos, está la asociación unidimensional y la asociación multidimensional; *asociación multinivel*, está basada en los niveles de abstracción que involucra, se divide en asociación de un nivel y asociación multinivel.

Las métricas que debemos tomar en cuenta para el uso de asociaciones son el *soporte*; es el número de veces con que A y B aparecen juntos en una base de datos, la *confianza*; es el cociente del soporte de la regla y el soporte del antecedente, y por último el *lift*; es el aumento de la probabilidad de que ocurra el consecuente como producto de que ocurrió el antecedente.

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, items o atributos que tienden a ocurrir de forma conjunta. *Apriori* fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, tiene dos etapas: la primera identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite y la segunda convertir esos itemsets frecuentes en reglas de asociación.

Detección de outliers

Los *outliers* o valores atípicos son aquellas observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos, sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar. Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

Existen cuatro categorías de outliers, los que surgen por un *error de procedimiento*, los que surgen debido a un *suceso extraordinario*, observaciones cuyos valores caen dentro del rango de las variables observadas pero que son *únicas en la combinación de los valores* de dichas y *sucesos extraordinarios para los que no se tiene explicación*.

Para la detección de outliers se debe examinar la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución, para ello se establece un umbral para la designación de caso atípico, esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas.

Visualización

La visualización de datos es la representación gráfica de información y datos, es el proceso de búsqueda, interpretación, contraste y comparación de datos que permite un conocimiento en profundidad y detalle de estos de tal forma que se transformen en información comprensible. Permite a los tomadores de decisiones ver la analítica presentada de forma visual, de modo que puedan captar conceptos difíciles o identificar nuevos patrones.

Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos, es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos. Las imágenes son emocionalmente atractivas, es más probable que las personas ofrezcan una reacción más intensa a una imagen que a un conjunto de palabras.

Existen muchas técnicas para la visualización, estas se basan según la complejidad y elaboración de la información. Se clasifican en 3 tipos: *elementos básicos de representación de datos*, algunos de estos son las gráficas, mapas y tablas; *cuadros de mando*, son composiciones complejas de visualizaciones individuales que guardan una coherencia y relación entre ellas; *infografías*, se utilizan para contar la historia, construyen narrativas a partir de los datos.

Las visualizaciones de datos hacen posible la narración de historias. Una visualización eficaz cuenta una historia, reduce la confusión que genera la acumulación de datos y resalta la información útil, una visualización de datos efectiva implica un equilibrio entre forma y función.

Existen al menos cuatro grupos de métodos para la visualización de datos, temporal: se refiere al tiempo y la investigación de cambios durante un período específico; jerárquico: representa una relación entre diferentes puntos de datos; de red: se trata de relaciones que se indican mediante líneas que conectan puntos y geoespacial: es una categoría que describe áreas geográficas e intenta transmitir una sensación de espacio.

Regresión

La regresión es de categoría predictiva, predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. Su objetivo es explotar la relación entre dos (o más) variables de modo que se pueda obtener información sobre una de ellas mediante el conocimiento de los valores de la otra u otras. Las regresiones se encargan de analizar la relación entre una variable dependiente y una o varias variables independientes, existe la *regresión lineal simple* y la *regresión lineal múltiple*, además de *la regresión polinomial*.

Cuando no se conoce la función conjunta de las variables (X, Y) y solo se cuenta con información muestral, la regresión se vuelve un problema de estimación de parámetros. Para una ecuación de regresión lineal simple, en la cual se considera al *componente determinístico* y al *error*.

El *análisis gráfico* de la relación entre las xi y las yi se hace mediante un *diagrama de dispersión*, según su forma se puede determinar si existe o no una relación lineal entre las variables.

El *método de mínimos cuadrados* es un enfoque estándar en el análisis de regresión para aproximar la solución de sistemas sobre determinados minimizando la suma de los cuadrados de los residuos hechos en los resultados de cada ecuación. La aplicación más importante es el ajuste de datos. Este método calcula a partir de los N pares de datos experimentales (x, y), los valores m y b que mejor ajustan los datos a una recta. Se entiende por el mejor ajuste aquella recta que hace mínimas las distancias d de los puntos medidos a la recta. Cuando se haga uso del método de mínimos cuadrados se debe buscar una línea de mejor ajuste que explique la posible relación entre una variable independiente y una variable dependiente.

Una de las principales aplicaciones del análisis de regresión es la proyección con diferentes escenarios. Esto, teniendo en cuenta el grado de influencia (en estadística se conoce a esto como correlación) sobre la variable dependiente.

Clasificación

La clasificación es la técnica más aplicada en la minería de datos, organiza o mapea un conjunto de atributos por clase, dependiendo de sus características. Empareja o asocia datos a grupos predefinidos, encuentra modelos que describen y distinguen clases o conceptos para futuras predicciones. Cada instancia pertenece a una clase distinguida por un tipo de atributo. Los demás atributos de la instancia se utilizan para predecir la clase de nuevas instancias. Esto se logra gracias a que se entrena un modelo usando datos recolectados para hacer predicciones futuras.

Algunas de las técnicas de clasificación son, *clasificación por inducción de árbol de decisión*, *clasificación Bayesiana*, *redes neuronales*, *Support Vector Machines (SVM)* y *clasificación basada en asociaciones*.

Los *Árboles de decisión d*efinen un conjunto de clases, asignando a cada dato de entra da una clase y determina la probabilidad de que ese registro pertenezca a la clase, a partir de una base de datos en donde se construyen los diagramas de construcciones lógicas, las *Redes Neuronales s*on modelos predictivos no lineales que aprenden a través del entrenamiento. Existen diferentes tipos de redes neuronales, las más conocidas son las simples y multicapas. Las tareas básicas de las redes neuronales son reconocer, clasificar, agrupar, asociar, almacenar patrones, aproximación de funciones, sistemas (predicción, control, entre otros) y optimizan. En términos simples, un clasificador de *Naive Bayes* asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable.

Patrones secuenciales

Los patrones secuenciales 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, se trata de

buscar asociaciones de forma que si sucede un evento en consecuencia sucederá otro en cierto lapso.

El *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 que expresa patrones de comportamiento secuencial, es decir que se dan en distintos momentos. Los patrones secuenciales se caracterizan porque el orden es importante.

Una *secuencia* es una lista ordenada de itemsets, donde cada *itemset* es un elemento de la secuencia, el *soporte* de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias, las *secuencias frecuentes* son las subsecuencias de una secuencia que tienen un soporte mínimo. La *agrupación de patrones secuenciales* es la tarea de separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a los objetivos de otros grupos. Para la creación de agrupamientos, se selecciona arbitrariamente el centro del primer agrupamiento, posteriormente, se procesan secuencialmente los demás patrones mediante cálculos de distancia. Cada M patrones se mezclan agrupamientos, pueden ser: *mezcla por cercanía*, *mezcla por tamaño* y *mezcla forzada*.

Las *reglas de asociación* con datos secuenciales se presentan cuando los datos contiguos presentan algún tipo de relación, estos expresan patrones de comportamiento secuenciales. Algunos métodos representativos son GSP, SPADE, AprioriAll, FreeSpan, SPAM, PrefixSpan, ISM, IncSp, ISE, IncSpan.

Predicción

La predicción tiene por finalidad obtener estimaciones o *pronósticos* de valores futuros de una serie temporal a partir de la información histórica contenida en la serie observada hasta el momento actual. Para realizar una predicción se necesita definir adecuadamente nuestro *problema*, *objetivos* y *salidas deseadas*, debemos *recopilar datos*, elegir una medida o indicador de éxito y preparar los datos que utilizaremos, debemos dividirlos en 70% *conjunto de entrenamiento*, 15% *conjunto de validación* y 15% *conjunto de pruebas*.

Un *árbol de decisión* es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente, facilita la toma de mejores decisiones, especialmente cuando existen riesgos, costos, beneficios, etcétera. Están formados por nodos y su lectura se realiza de arriba hacia abajo, dentro de un árbol de decisión encontramos diferentes tipos de nodos, *nodo raíz*, *nodos intermedios* y *nodos terminales*.

Para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones. Si una subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos en la misma clase. Los árboles se pueden clasificar en dos tipos, *árboles de regresión* en los cuales la variable respuesta es cuantitativa; *árboles de clasificación* en los cuales la variable respuesta es cualitativa.

Un *árbol de clasificación* consiste en hacer preguntas para las covariables cuantitativas y cualitativas, aplica la estrategia "divide y vencerás" para hacer la clasificación, implementando métodos y técnicas para la realización de procesos inteligentes, representando así el conocimiento y el aprendizaje, con el propósito de automatizar tareas.

Un *árbol de regresión* consiste en hacer preguntas $x_k \leq c$? para cada una de las covariables, son un método de analítica de datos y se usan cuando queremos predecir el valor de una variable numérica, predice valores de respuestas mediante el aprendizaje de reglas de decisión derivadas de características.