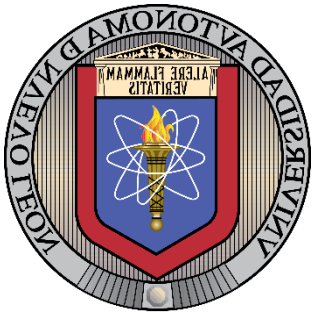

UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

FACUTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Minería de Datos

Resumen de Técnicas de Minería de Datos

Maestra: Mayra Cristina Berrones Reyes

Alumno: Juan Alfredo Cantú Zavala

Matrícula: 1810736

Reglas de asociación

Es la 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, básicamente busca patrones en ciertos conjuntos para poder realizar “predicciones” en base a ciertos patrones repetitivos que se infiere que se pueden repetir. Su aplicación es muy variada va desde análisis de la banca tales como prestamos bancarios, en el cual podemos generar ciertas “reglas” que nos permiten administrar de mejor manera los riesgos en tema de préstamos, en inversiones de la propia empresa, etc. También tiene aplicaciones en temas de Cross-Marketing que es básicamente artículos que uno como consumidor usualmente se compramos juntos como por ejemplo cuando se compra un refresco casi siempre se compran alguna botana o alimento y como ese hay muchos más ejemplos, otra de las aplicaciones de estas reglas de asociación es en el diseño de los catálogos y en el acomodo de los artículos de una tienda

Los ejercicios de este tema se realizan contando las veces en que un par o más artículos se adquieren en la misma compra, una vez teniendo esta cifra se divide entre el total de compras realizadas en el periodo de tiempo analizado y con el cociente de esta división obtenemos la confianza, y esa es solo una de las técnicas, podemos mencionar otras basadas principalmente en diagramas de árbol o un diagrama de red y se van analizando cada nivel y se usan una especie de filtros para ver si realmente cumplen con cierto nivel o numero de ocurrencias que al dividirse entre el total de experimentos se obtiene un número llamado soporte en el cual mientras más alto sea este número es mas probable que estemos hablando de una regla de asociación que obtuvimos utilizando alguno de los algoritmos.

Regresión

Una regresión es un modelo matemático para determinar el grado de dependencia entre uno o mas variables, es decir se busca saber conocer si existe relación entre ellas existen dos tipos de regresión la regresión lineal y la regresión lineal múltiple, en la primera se analiza el caso en que una variable independiente ejerce o no influencia sobre otra variable dependiente tenemos, en otras palabras se tienen un conjunto de datos en un plano y se estima la ecuación de una recta que trate de tocar o estar lo mas cerca posible de todos los datos los datos graficados en el plano, existen dos herramientas que nos permiten saber que tan bien ajusta nuestro modelo a las datos primero tenemos la prueba de significancia que es una prueba de hipótesis en las cual se busca ver si realmente existe una relación entre la variable independiente y la dependiente y la segunda herramienta es la R ajustado que una vez probado la significancia del modelo nos

indica que tan bien ajusta nuestra recta a los datos por lo que lo que buscamos es acercarnos lo mas posible a un 100% de ajuste, por otra parte la regresión lineal múltiple es el mismo proceso que el modelo lineal con la diferencia de que en este se tienen dos o más variables independientes, se tiene las mismas dos herramientas que con el modelo lineal tenemos agregamos una mas que es la R cuadrada ajustada la cual es un valor que también buscamos que resulte cerca del 100% con la diferencia de que este numero nos dice si vale la pena agregar una variable dependiente, es decir si tenemos tres variables independientes y obtenemos una R cuadrada ajustada de 99% y si al agregar una cuarta variable este porcentaje no aumenta significa que ni agregando mas variables independientes el ajuste del modelo va a ser mejor por lo que el modelo optimo será el que tenga una R cuadrada ajustada más cercana al 100%, por lo que si al tener dos variables independientes y realizar el proceso con una tercera si el valor de R cuadrada ajustada no aumenta podemos inferir que el modelo más optimo es el modelo de dos variables.

Clasificación

La clasificación es una técnica utilizada en la minería de datos que consiste en el ordenamiento o disposición de datos por clases tomando en cuenta ciertas características de estos datos, algunos de los métodos que se utilizan en esta técnica son el Análisis discriminante, que es un método utilizado para encontrar una combinación lineal de rasgos que separan las clases de objetos o eventos, dicho de otra manera es verificar ciertos rasgos distintivos de los datos para así agruparlos basados en la característica comparada, también están los Árboles de decisión que son un método analítico en el que en cada nodo tenemos una decisión o el resultado de un experimento y en cada rama tenemos un resultado y dependiendo de la decisión o resultado del experimento tomamos otro nodo que nos lleva a otra rama y con esto podemos analizar los distintos escenarios en los que nos podríamos encontrar según nuestras decisiones o resultados obtenidos, otro método son las Reglas de clasificación estas buscan términos no clasificados periódicamente o no muy frecuentes y en caso de encontrar una coincidencia entre estos datos los clasifica junto a los otros que si se encuentran periódicamente, finalmente tenemos las Redes neuronales artificiales es un modelos de unidades conectadas que transmiten señales, entre otros. Cada uno de estos métodos tienen ciertas características como los son la precisión de la predicción que como su nombre lo indica es cuanto se acerca nuestra predicción a nuestros valores reales, la eficiencia que es cuentos recursos consumen considerando el tiempo y la exactitud, la robustez que es cuan grandes son estos algoritmos, la escalabilidad en donde vemos las escalas utilizadas y las aplicaciones y la interpretabilidad que es que tan fácil de entender los datos obtenidos que nos arroja cada método. Un ejemplo muy sencillo de clasificación es cuando vemos que normalmente en Estados unidos las calificaciones que los maestros asignan son A si la

calificación esta entre 90 y 100, B si la calificación esta entre 80 y 90 (sin incluirlo), C si esta esta entre 70 y 80 (sin incluirlo), D si tiene una calificación entre 60 y 70 (sin incluirlo) y F si la calificación es menor de 60, entonces si un maestro decidiera calificar a sus alumnos con valores numéricos dependiendo del número podemos asignar una letra según el rango en que cada calificación se encuentre.

Outliers

Los outliers también conocidos como valores atípicos son valores que muestran un comportamiento extremo que difieren del patrón general de una muestra. Los datos atípicos son generados ocasionalmente por errores por datos de entrada y procedimiento, acontecimientos extraordinarios, valores muy extremos y o faltantes, o bien causas no conocidas, estos resultados atípicos distorsionan los resultados de los análisis, y por esta razón es necesario identificarlas y tratarlos de manera adecuada, ahora ¿Cómo podemos calcular estos valores atípicos? Vamos a dividir esta respuesta en dos categorías principales primero tenemos Métodos univariantes para la detección de outliers y los Métodos multivariantes para la detección de outliers, luego tenemos algunas de las técnicas para la detección de los valores atípicos tales como la Prueba de Grubbs también conocida como prueba residual máxima normalizada ente esta prueba se compara el valor máximo de la desviación con la media para después dividir todo entre la desviación, después la Prueba de Dixon, en esta prueba se compara la cantidad de valores extremos inferior y superior y en caso de haber pocos valores extremos podría tratarse de un valor atípico, luego tenemos la Prueba de Turkey en esta prueba tenemos una grafica de boxplot en donde hay un dato fuera de la caja y de los límites inferior y superior, tenemos también la técnica de Análisis de valores o Atípicos de Mahalanobis aquí se ve la distancia entre un punto P (generalmente la media) y la desviación estándar de cada dato y se presta atención en las distancias mas grandes y finalmente la Regresión simple que puede ser lineal, cuadrática o polinómica en la cual se busca ajustar ya sea una línea o una curva dependiendo de la situación para buscar predecir datos futuros pero si tenemos un modelo que ajusta bien a nuestros datos a la hora de ir a la grafica si hay un punto que se aleja mucho de la línea o curva estimada podría tratarse de un valor atípico, algunos de los softwares más utilizados son R, Excel, Minitab, Tableau, Google Analytics, entre otros, una vez detectados estos valores se debe verificar si se deben a un error de captura o en la medición de la variable ya que en caso decidir no tomar en cuenta los valores extremos podemos estar introduciendo sesgo al estudio, reduciendo el tamaño de nuestra muestra, puede afectar a la distribución y las varianzas, algunas de las aplicaciones que le podemos dar a la detección de estos datos son para la detección de fraudes financieros, en tecnología informática y telecomunicaciones, Nutrición y salud, Negocios, etc. Ya detectados los valores atípicos estos pueden tener distintos significados ya sea que sean errores lógicos o de captura de datos, error de limites en este los valores pueden esterse

escapando del grupo medio, pero a fin de cuentas mantenemos el dato para no modificar el estudio o errores de valores anómalos en que podría ser el caso en el que sean los valores que estemos buscando.

Patrones secuenciales

En Minería de datos 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 cabe destacar que el orden del acontecimiento es considerado, el objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos. Las reglas de la sucesión secuencial expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos de tiempo. Las principales características son que el orden importa, el objetivo es encontrar los patrones secuenciales, el tamaño de una secuencia es su cantidad de elementos, la longitud de la secuencia en la cantidad de ítems, el soporte de la secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S y las frecuencias frecuentes son las subsecuentes de una secuencia que tiene un soporte mínimo, entre otras, las ventajas que tiene este esta técnica son la flexibilidad ya que puede ser que el patrón varíe un poco pero no termine por cambiar el resultado y la eficiencia al identificar de una manera rápida los patrones y dependiendo del algoritmo consumiendo los recursos necesarios y la desventajas pueden encontrarse en si en la utilización porque este algoritmo puede ser tan complicado como lo hagamos y que puede ser sesgado por los primeros patrones los tipos de datos que pudiéramos manejar utilizando su técnica son el ADN y proteínas, el recorrido de los clientes en un supermercado, registros de un acceso a una página web, en las aplicaciones encontramos el campo de la medicina, al buscar predecir ciertos compuestos que causen enfermedades, el comportamiento de una persona en cuando realiza sus compras esto puede ir desde el recorrido que el cliente sigue en la tienda, los artículos que compra juntos, los artículos que mas se venden o el caso contrario que menos se venden, estos dos ejemplos los podemos entender como el agrupamiento de patrones secuenciales y otro ejemplo podría ser el reconocimiento de spam de un correo electrónico, es te reconocimiento es posible ya que el usuario reporta ciertos correos como spam y lo que hace la computadora es identificar las palabras clave para poder darse cuenta en un futuro si se trata de un correo con spam, este ejemplo comprende la clasificación de datos secuenciales.

Predicción

La producción 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 condiciones relativas a la relación temporal de los variables entrada o proyectores de la variable objetivo que son los valores generalmente son continuos y las predictores de la variable de respuesta, esta técnica se utiliza para datos que a menudo son del futuro, tenemos también variables independientes que son los atributos que ya conocemos y las variables de respuesta que básicamente son lo que queremos saber, está relacionado con otras técnicas porque cualquiera de las técnicas utilizadas para la clasificación estimación puede ser adaptado 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 utilizan para construir un modelo que explique el comportamiento observado en los datos cuando, en este modelo se aplican nuevas entradas de datos, el resultado es una producción del comportamiento futuro de los mismos, las aplicaciones de esto van desde revisar los historiales crediticio de los consumidores y de las compras pasadas para predecir si serán un o representaran un riesgo crediticio en el futuro para la compañía o la persona que decida prestarles o bien predecir el precio de venta de una propiedad dentro de unos años, predecir si va a llover en función de la humedad actual, predecir la puntuación de cualquier equipo de fútbol durante un partido, entre otros, las técnicas que se utilizan se va a hacer en modelos matemáticos generalmente en modelos estadísticos simples como la regresión, estadísticas no lineales como series de potencias, o bien redes neuronales, todo esto basado en ajustar una curva a través de los datos que nos permita explicar de una buena manera el comportamiento de nuestros datos, en otras palabras, encontrar una relación entre los productores y los valores que queremos pronosticar.

Visualización de datos

Nosotros que toda la vida hemos tenido algún conjunto de datos a nuestro alcance sólo que con el tiempo nos hemos hecho capaces de aplicar en ellos ciertas herramientas o técnicas que nos permiten no sólo interpretar esos datos, sino que también darle un uso para nuestro beneficio o sacarle provecho a los mismos. Las técnicas en la visualización de datos que nos sirven para representar gráficamente los elementos más importantes en una base de datos, sabiendo esto nos planteamos lo siguiente ¿qué es la visualización de datos? La visualización de datos es la presentación de información en el 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

datos, existen diferentes tipos de visualización de datos ya que uno de los grandes retos que enfrentan los usuarios de empresas, es qué tipo de elemento visual utilizar para representar la información de mejor forma. Aunque existen muchos tipos algunos de los más comunes son los gráficos, este tipo es el más común y conocido que utilizamos en nuestro día con hojas de cálculo como para representar datos de manera sencilla, como datos circulares, líneas columnas y barras, burbujas, áreas, diagramas de dispersión y mapas de tipo árbol otro tipo o forma de visualización, también tenemos los mapas como una forma de visualizar datos geográficos, todos alguna vez hemos realizado una visualización de datos en mapas en donde básicamente vemos un dibujo en segunda dimensión de cierto lugar del cual queremos localizar alguna dirección, también tenemos lo que son las infografías estas son una colección de imágenes, gráficos y textos simples que resume un tema para que se pueda entender más fácilmente este recurso resulte excelente para ayudarnos a procesar más fácil la información más compleja también tenemos los cuadros de mando o Dashboard qué básicamente son un cuadro de mando que se utiliza como herramienta que nos permite saber en todo momento el estado o indicadores de nuestro negocio.

La mayoría de los analistas de datos utilizan software amansado por explorar visualizar los datos y las herramientas de software te mande es de hojas de cálculo sencillos en Excel o Google Sheets a software de análisis más sofisticado, En su aplicaciones se encuentran básicamente cualquier usuario o persona o empresa que quiero conocer sus condiciones de una manera fácil de digerir utilizado básicamente en cualquier área empresarial ya que permite digerir de una manera muy fácil grandes cantidades de datos que pueden llegar a ser complejos.

Clustering

El clustering o bien el agrupamiento es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos o con ciertas características similares, las técnicas de clustering se obtienen utilizando algoritmos matemáticos que se encargan de agrupar objetos. Puede ser usando la información que brindan las variables que pertenecen a cada grupo con esta información los algoritmos miden la similitud entre los mismos, una vez que se hace esta medición se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las otras clases un cluster es una colección de objetos de datos muy similares entre sí dentro del mismo grupo y diferente a los objetos de otros grupos, supongamos que tenemos un conjunto de datos y si queremos entender su estructura tendremos que entender las similitudes entre los datos con las características encontradas en ellos mismos esto es un aprendizaje no supervisado ya que no hay clases predefinidas el clustering tiene aplicaciones en aseguradoras cuando queremos identificar los grupos con alto costo promedio de reclamo, también se utiliza

para el uso de suelo con esto se puede identificar áreas de uso similar de la tierra en base a datos de observación de la tierra ya sea un área comercial o apta para la agricultura, o bien se utiliza en el marketing en donde permite ayudar a los profesionales a descubrir distintos grupos de bases de clientes, en la planeación de una ciudad identificando grupos de casas según su tipo valor ubicación y cualidades geográficas, lo podemos ver en ciertos estudios de terremotos los epicentros del terremoto observados deben agruparse a lo largo de las fallas continentales, existen diferentes métodos de agrupación tenemos asignación Jerárquica frente a un punto, Datos Numéricos y/o Simbólicos, Determinística vs. Probabilística, Exclusivo vs. Superpuesto, Jerárquico vs. Plano y De Arriba a Abajo y De Abajo a Arriba, entre otras existen varios algoritmos del clustering y entre ellos encontramos el K-Means y él X-Means, que son algoritmos parecidos con ligeras variaciones que se usan mucho para esta tecnica.