UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias físico Matemáticas

Minería de Datos

M.C. Mayra Cristina Berrones Reyes

Resúmenes de técnicas de minería de datos.

Gpo 003

Alumno
Manuel Joseph Romero Pascacio 1811177

INDICE

Contenido

Reglas de Asociación.

Las reglas de asociación es una técnica que se utiliza en la inteligencia artificial en el data meeting lo que hace es describir una regla como su nombre lo indica de asociación entre los conjuntos de datos relevantes. Es la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras entre conjuntos de elementos.

Los conceptos que se utilizan en las reglas de asociación son los siguientes:

• Conjunto de elementos: es la colección de uno o más artículos
• Recuento de soporte: es la frecuencia de ocurrencia de un ítemset.
• Confianza: mide que tan frecuentes son los ítems y aparecen en transacciones, que se puede ver como una probabilidad condicional entre los datos y su asociación

Reglas de Apriori
Para las reglas de asociación utilizamos el Principio A priori que dice que, si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben de ser frecuentes.

Clasificación.

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

Métodos de clasificación:

- 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

Las características finales que podemos ver en clasificación son:

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

Outliers.

La detección de outliers estudia es comportamiento de valores extremos que difieren del patrón general de una muestra muchas veces estos son conocidos como los valores atípicos dentro de un conjunto de datos, porque son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos, por lo que distorsionan los resultados de los análisis y por esta razón hay que identificarlos y tratarlos de manera adecuada.

Si no es un error, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esa información, debido a que se introducen un sesgo, disminuye el tamaño muestra y/o puede afectar la distribución y varianzas. Por lo tanto, la mejor opción es quitarles peso a esas observaciones atípicas mediante técnicas robustas

Aplicación de la minería de datos en outliers:

- Detección de fraudes financieros
- Tecnología informática y telecomunicaciones
- Nutrición y salud
- Negocios

Patrones secuenciales.

En minería de datos secuenciales, los cuales es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el tiempo. Reglas de asociación secuencial representan patrones en distintos lapsos del tiempo.

Características

- El orden importa.
- El objetivo es 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 las secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene un soporte mínimo.

Ventaja:

- Flexibilidad
- Eficiencia

Desventajas:

- Utilización: es prueba y error
- Sesgado por los primeros patrones

Aplicaciones

- Agrupamiento de patrones secuenciales: objetos de un grupo similares se agrupa para las características entre sí y diferentes a otros grupos.
- Clasificación con datos secuenciales: funciona como un algoritmo de patrones que se repiten.

Predicción.

Técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento. En algunos casos, el simple hecho de conocer y comprender las tendencias históricas es suficiente para trazar una predicción 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 y como se mencionó anteriormente, las predicciones son a menudo sobre el futuro.

Aplicaciones:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Predecir el precio de venta de una propiedad.
- Predecir si va a llover en función de la humedad actual.
- Predecir la puntuación de cualquier equipo durante un partido de fútbol

Técnicas de predicción:

- 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.

Tipos de regresiones:

- Regresión lineal: cuando una variable independiente ejerce
- Regresión lineal múltiple

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

- Variables dependientes: Factor el cual se está tratando de entender o predecir
- Variables independientes: Factor que se cree que puede impactar en la variable dependiente

Visualización.

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

- Gráficos
- Mapas
- Infografías
- Cuadros de mando

Las aplicaciones que podemos entender de la visualización de datos son:

- 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.

Clustering

Cuando hablamos de Clustering nos referimos a la colección de objetos de datos, esto consiste en la división de los datos en grupos de objetos similares, usando la información que nos brindan las variables.

Estos tienen que ser similares dentro del mismo grupo mas no iguales. Una vez obtenidos los grupos clúster podemos graficarlos, obteniendo una grafica de puntos, con lo cual podemos hacer un análisis. El análisis de cluster dado la grafica de puntos es para entender la estructura, y encontrar similitudes entre los datos de acuerdo a las características encontradas

Métodos de Agrupación:

- Asignación jerárquica frente a un punto
- Datos numéricos y/o simbólicos
- Determinística vs probabilidad
- Exclusivo vs superpuesto
- Jerárquico vs plano
- De arriba abajo y de abajo a arriba

Algoritmos del cluster:

- Simple K-Means
- X-Means
- EM
- Cobweb