UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

FCFM

Minería de Datos

Resumen de técnicas de Minería de Datos

Patricia García Olmeda

1931549

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

Reglas de asociación:

Las reglas de asociación buscan patrones frecuentes, asociaciones, correlaciones o estructuras entre conjuntos de elementos u objetos en bases de datos.

Esta 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.

Los conceptos utilizados son:

- Conjunto de elementos
- Recuento de soporte
- Confianza

Reglas de Asociación en la Minería:

1. Generación de elementos frecuentes: generar todos los conjuntos de elementos con soporte mínimo superior
2. Generación de reglas: generas las reglas de alta confianza a partir de un conjunto de elementos frecuentes, cada regla es una partición binaria de un conjuntos de elementos

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

Características de los métodos:

- Precisión de 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

Se puede decir que son los valores atípicos en 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 ay que identificarlos y tratarlos de manera adecuada

Se pueden eliminar o sustituir los outliers si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable.

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:

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 secuencia 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

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 menciono anteriormente, las predicciones son a menudo sobre le futuro.

Aplicaciones:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si son una opción rentable en el futuro.
- Predecir el precio de venta de una propiedad
- Predecir si lloverá en función de la humedad actual
- Predecir la puntuación de cualquier

Regresión:

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o mas 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 mas variables e identificar cuales 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 esta tratando de entender o predecir
- Variables independientes: Factor que se cree que puede impactar en la variable dependiente

Visualización de datos:

Nos sirve para representar gráficamente los elementos mas importantes de nuestra base de datos. La visualización de datos es la presentación de información en un formato ilustrado o grafico. Al utilizar elementos visuales como cuadros, gráficos o mapa, 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

Aplicaciones

- Identificar relaciones y patrones
- Comprender la información con rapidez
- Identificar tendencias emergentes
- Comunicar la historia a otras personas