Análisis de Grandes Volúmenes de Datos
Carlos Espinoza Peraza - B92786
Reporte 24 de mayo: Clustering

Resumen de los temas estudiados

Análisis Asociativo

Es el estudio de atributos o características que se suelen dar en conjunto. Esto se hace mediante métodos que prueban las relaciones que hay entre los atributos.

> Usos que se les da: Compras, Detección de Fraudes, Análisis del Mercado y Clientes.

¿En qué se basa?

- Utilizan datos transaccionales
- Utiliza la terminología de `itemset` que es el conjunto de items
- Número de items en la trasacción
- `transaction width` se refiere a la cantidad de items que tiene una transacción
- `frecuencia` es el número de transacciones que contienen el itemset

Reglas

- Reglas de forma: X -> Y donde no exista intersección
- Propiedades de Soporte: Para distinguir un evento que ocurre por casualidad
- Propiedades de Confianza: Medida de precisión. Cuan mayor sea es más probable que un atributo aparezca en X transacción
- Definir granularidad y prioridades

Generar Reglas

- Encontrar los itemsets frecuentes
- Generar reglas de asociación

Aspectos a tomar en cuenta

- No todas las reglas son útiles
    - Actionable rules: Deben agregar valor al negocio
    - Trivial rules: Reflejan asociaciones lógicas
    - Inexplicable rules: podrían no obedecer a ninguna razón clara

-Lift: proporciona una medida de utilidad de la regla

Patrones locales y modelos

- **Modelo:** da una descripción general del dataset
- **Patrones**: ocurren en sectores de los datos
- **Técnicas de asociación:** encuentran patrones en su mayoría locales

Series de Tiempo

En algunos casos el atributo label es tomando en cuenta en su relación con el tiempo, por lo que se da un conjunto de observaciones dada una dimensión temporal.

Interés

- Identificar patrones
- Hacer predicciones

Las series de tiempo son secuencias de valores que representan una variable de interés, pueden ser:

- Discretas
- Continuas

Componentes

- **Tendencia:** Describe el comportamiento promedio de la serie
- **Temporalidad:** Olas de frecuencia regular
- **Ruido:** Tipo de fluctuación que representa variaciones irregulares en los datos
- **Descomposición de una serie de tiempo:** Consiste identificar los tres elementos (tendencia, temporalidad y ruido)
- **Moving average: **

Eliminación de estacionalidad

- **Diferenciación**
- **Tests estadísticos**

Exponential smoothing model (ES)

- **ES simple**
- **ES con ajuste de tendencia y temporalidad**
- **ES adaptativo**

Autoregressive Models

- **Auto regresión ACF y PACF**
- **Moving average MA y AR moving average**
- **AR integrated moving average ARIMA**

- **Seasonal AR Integrated moving average SARIMA**

Comentarios sobre la materia estudiada

El uso del análsis de asociatividad permite, al igual que los clusters, tener un mayor entendimiento de los datos con los que se está trabajando. Se pueden utilizar en una amplia variedad de problemas y son funcionales en la medida que aportan un valor real al negocio. El conocimiento de estas técnicas puede influir directamente en la competitividad de los análistas de datos.

Dudas sobre la materia

> Ninguna

Posible uso como profesional

En casos en los que se necesite hacer un análsis de tendencia a lo largo del tiempo, se podría usar la serie de tiempo para analizar los valores cambiantes en un periodo determinado.

Problemas que podría resolver con las técnicas estudiadas

Para un problema en el que se necesite analizar cuales son los factores más influyentes a la hora de realizar una clasificación, se podría determinar los atributos de mayor peso que logran la determinación de un label. Ejemplo, que factor es reincidinte en los casos de fraude.