UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma De Nuevo Leon

Monterrey, Nuevo Leon

Alumno

Rafael Hernan Elizondo Aranda

1887942

Materia

Minería de Datos

Maestra

Mayra Berrones

1. **Clasificación:**

La clasificación es la técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características, consiste en adaptar un modelo para hacer predicciones futuras.

El método de clasificación consiste en agrupar los elementos de un conjunto de datos por sus características similares, esto con la finalidad de poder predecir a futuro nuevos elementos que se introduzcan o conocer las características de los elementos ya existentes en los datos, existen diversos métodos para aplicar la clasificación

- **Clasificación por inducción de árbol de decisión:** Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos. • Útiles en Clasificación, Agrupamiento, Regresión
- **Clasificación Bayesiana** Si tenemos una hipótesis H sustentada para una evidencia E -> p(H|E)= (p(E|H)* p(H))/p(E) Donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B, este método se utiliza para predecir los eventos futuros conociendo las probabilidades individuales de las clasificaciones.
- **Redes neuronales** Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse, Se usan en Clasificación, Agrupamiento, Regresión Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida
- **Support Vector Machines (SVM)**, **Clasificación basada en asociaciones**

La técnica de clasificación puede ser muy útil para trabajar, pero también tiene limitantes y hay que saber con que datos es mejor trabajar, los datos nominales son más fáciles de trabajar que por ejemplo datos descriptivos.

2. **Visualización**:

La visualización es una técnica descriptiva y consiste en la representación grafica de los elementos de una base de datos

La visualización de datos y, en concreto, el uso de las herramientas que cumplen este cometido es fundamental para agilizar el proceso y ahorrar tiempo y esfuerzos a los expertos que deben determinar, con la máxima rapidez y eficiencia, si los modelos obtenidos se ajustan con lo esperado. En esa comparativa entre modelos y su evaluación para determinar si son lo suficientemente satisfactorios es donde entran en juego las herramientas de visualización de datos, que simplifican y agilizan la tarea de los expertos permitiendo optimizar el proceso de minería de datos, reduciendo el tiempo empleado para llevarlo a cabo y minimizando los riesgos asociados a una mala interpretación de los resultados obtenidos.

Las elementos mas usados en la visualización de datos son :

- **Gráficos**
- **Mapas**
- **Tablas**
- **Cuadros de Mando** Un Cuadro de Mando está configurado por KPIs acompañados de una representación gráfica**,** de esta forma se puede acceder a la información de manera muy visual y ágilmente. Este tipo de herramienta permite optimizar los procesos de toma de decisiones tanto estratégicas como tácticas.
- **Infografias** Una infografía es una imagen explicativa que combina texto, ilustración y diseño, cuyo propósito es sintetizar información de cierta complejidad e importancia, de una manera directa y rápida. Las infografías responden a diferentes modelos, tales como diagramas, esquemas, mapas conceptuales, entre otros. Asimismo, utilizan diferentes tipos de recursos visuales, tanto lingüísticos como no lingüísticos: texto, imágenes, colores, criterios de diagramación y toda suerte de elementos plásticos y compositivos que sean de provecho.

3. Patrones Secuenciales

Consiste en encontrar patrones estadísticamente relevantes en colecciones de datos que están representados de forma secuencial. Debido a la frecuencia con que aparecen este tipo de datos en escenarios de aplicaciones reales, esta técnica constituye uno de los métodos más populares de descubrimiento de patrones.

Los patrones secuenciales como lo dice el nombre buscan patrones que seguir, por ejemplo cuando un cliente compra un celular es muy probable que compre un protector de pantalla, eso es lo que busca los patrones secuenciales buscar las tendencias de los clientes.

El método para resolver por patrones secuenciales consiste en tres pasos :

1 agrupamiento de los patrones

2.Clasificar los patrones

3, Definir las reglas de asociación de los patrones secuenciales

Los Métodos representativos son :

- GSP [Generalized Generalized Sequential Sequential Patterns Patterns].
- SPADE [Sequential PAttern Discovery using Equivalent Class]
- CloSpan
- Patrones secuenciales con restricciones.
- SSMiner SSMiner [Similar [Similar Sequence Sequence Miner]

4. **Regresión:**

Es un proceso estadístico para estimar las relaciones entre variables. Incluye muchas técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes (o predictoras). Más específicamente, el análisis de regresión ayuda a entender cómo el valor de la variable dependiente varía al cambiar el valor de una de las variables independientes, manteniendo el valor de las otras variables independientes fijas.

Sirve para predecir como se va a comportar la variable independiente bajo determinadas condiciones, ya sea tiempo, cantidad etc… para este método puede existir una sola variable y seria regresión lineal simple

Y=B1*X+Bo+E

Un método para resolver esta ecuación es el método de mínimos cuadrados y consiste en la siguiente manera

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

Cuando se tienen mas variables que nos dicen como se comporta nuestra variable dependiente entonces tenemos una regresión lineal múltiple que se comporta así:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

La condición de este modelo es que si tenemos K variables predictoras entonces necesitamos k+1 para poder obtener la ecuación de la regresión lineal múltiple.

De esta manera se le puede dar uso a la regresión lineal, que es útil en distintos ámbitos como medicina, ventas, producción y muchos otros

5. **Outliers:**

Es una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. Por ejemplo, en el cálculo de la temperatura media de 10 objetos en una habitación, si la mayoría tienen entre 20 y 25 ºC, pero hay un horno a 350 °C, la mediana de los datos puede ser 23, pero la temperatura media será 55. En este caso, la mediana refleja mejor la temperatura de la muestra al azar de un objeto que la media. Los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de las muestras establecidas

Después de detectar los valores atípicos no es necesario eliminarlos de la base de datos o omitirlos puesto que al hacer esto se crea un sesgo que se busca emitir, lo que su busca hacer es quitarle peso en el resultado final mediante técnicas robustos, que son técnicas que se ven menos afectadas por grandes variaciones.

El encontrar valores atípicos es necesario cerciorarse de que no se deba a un error primero que nada. Esta técnica es muy útil en las aseguradoras o en los bancos para evitar fraudes pues ayuda a detectar cuando los números no dan y evitar robos.

6. Predicción

Como lo dice el nombre es un modelo predictivo, antes de poder utilizar este método es importante poder definir nuestro problema de e identificar las variables de salida, después se recopilaran los datos, elegir un indicador de éxito y por ultimo preparar los datos. Es necesario dividir los datos de manera eficiente 70% se irán al conjunto de entrenamiento, 15% al conjunto de validación y 15% al conjunto de pruebas.

Para este modelo se utilizan arboles de decisión

Árbol de regresión: cuando la variable y es cuantitativa

Árbol de clasificación : cuando la variable y es cualitativa

Los arboles están formados por nodos los cuales se leen de arriba abajo y los nodos están conformados de la siguiente manera

Primer nodo: es la primera división en función de la variable mas importante

Nodos intermedios: tras la primera División sigue esta división en base de estas variables

Nodos terminales: se ubican en la parte inferior y su función es definir la clasificacion definitiva

Ahora los existen dos tipos de nodos los de decisión que tienen una condición al principio y mas nodos debajo de ellos y también existen los nodos de predicción que no tienen condiciones ni nodos debajo de ellos.

Un técnica utilizada son los bosques aleatorios que consisten en una técnica de aprendizaje automático basada en arboles de decisión su principal ventaja es que tiene mejor rendimiento el cual consiste en la mejora de compensación de los árboles de decisión, y para asegurarnos de que cada árbol sea distinto, se enfrenta cada uno de ellos con una muestra aleatoria de los datos de entrenamiento.

7. Clustering

Es una técnica de aprendizaje no supervisada que consiste en agrupar puntos de datos y de esta forma crear similitudes

Existen cuatro tipos de análisis

- **Centroid Based cluster** se agrupan en conjuntos llamados centroides, los clusters se construyen basados en la distancio de los puntos hasta el centroide. El algoritmo mas común para esto es el de K medias

- **Connectivity Based Cluster:** Los cluster se definen agrupando los datos mas similares o cercanos, mientras mas cerca mas similar, un cluster contiene mas clusters

- **Distribution Based Clusters:** cada cluster pertenece a una distribución normal, la idea es que los puntos sean divididos en base a la probabilidad

- **Density Based Clusters:** Los clusters son definidos por áreas de concentración, se trata de conectar puntos cuya distancia se considera pequeña.

El método de K medias es un algoritmo de clustering basado en centroides donde k representa el numero de clusters y es definido por el usuario. Después el primer paso es seleccionar k datos aleatorios que serán los centroides segundos se analizara la distancia de cada uno al centroide más cercano y así obtener la media y será el nuevo cluster, después repetimos el proceso hasta que no existan cambios en la gráfica. Mientras aumentemos el valor de k la varianza disminuye y entre menor sea la suma de las varianzas mejor es el cluster.

8. Reglas de Asociación:

Las reglas de asociación se extraen de un tipo de análisis que se extraen información por coincidencias con el objetivo de encontrar relaciones en las transacciones. Un conjunto de asocianes se define como si primero va A entonces después debe Ir B por ejemplo ,si primero tenemos leche después debe ir cereal. Las reglas de asociación nos permiten encontrar los artículos que se asocian con mayor frecuencia y medir la fuerza de estas combinaciones

Tenemos dos tipos de Asociación

Asociación Cuantitativa: con base en los tipos de valores que manejan las reglas

Asociación Multidimensional: con base en las dimensiones de datos que involucra la regla

Dentro de las Asocianes existen tres tipos de métricas

Soporte: nos indican con que frecuencia aparecen juntos dos ítems en una base de datos

$$Soporte(A \Rightarrow B) = P(A \cap B)$$

Confianza: el el cociente de la probabilidad de la regla, es la probabilidad condicional

$$Confianza(A \Rightarrow B) = \frac{Soporte(A \Rightarrow B)}{Soporte(A)} = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

Lift: refleja el aumento de una probabilidad de que ocurra el consecuente cuando sabemos que ya ocurrió el anterior.

$$Lift(A \Rightarrow B) = \frac{Soporte(A \Rightarrow B)}{Soporte(A)Soporte(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$