Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de Datos

Resúmenes

Técnicas de Minería de Datos

Helena Patricia Carrillo Soto

Grupo: 002 Matrícula: 1725370

A 01 de Octubre de 2020, San Nicolás de los Garza, Nuevo León

Técnicas de Minería de Datos

Descriptivas

El objetivo de este tipo de técnicas es encontrar patrones que den un resumen de las relaciones ocultas dentro de los datos. Nos ayudan a descubrir las características más importantes de la base de datos.

Clustering

El clustering es una técnica de aprendizaje de máquina no supervisada (no se tiene una clase de respuesta) que consiste en agrupar puntos de datos y de esta forma crear particiones o clústers basándonos en similitudes. Se usa para investigaciones de mercado, identificar comunidades, prevención de crimen, procesamiento de imágenes, entre otros.

En clustering algunos datos tienen que estandarizarse en caso de ser cuantitativos y binarizarse en caso de ser de tipo categórico.

Existen 4 tipos de básicos de clustering:

- ***Centroid Based Clustering*:** Cada clúster es representado por un centroide. Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más es el de K-medias.
- ***Connectivity Based Clustering**:* Los clusters se definen agrupando a los datos más similares. La característica principal es que un clúster contiene a otros clusters, y debido a esto representan una jerarquía. Funciona de dos maneras: ascendente (de clústers más pequeños a más grandes) o descendente (de más grandes a más pequeños). Se representan por un dendrograma que muestra la jerarquía. *Hierarchical clustering* es un algoritmo de perteneciente a este tipo.
- ***Distribution Based Clustering:*** Cada clúster pertenece a una distribución normal, los puntos son divididos basados en la probabilidad de pertenecer a la misma distribución normal. Se tienen probabilidades de pertenecer a un grupo en lugar de asegurar que se pertenece de manera definitiva lo permite un mejor manejo de datos atípicos. Un algoritmo perteneciente a este tipo es *Gaussian mixture models*.
- ***Density Based Clustering:*** Los clústers son definidos por áreas de concentración. Se buscan áreas de puntos concentrados y se asignan esas áreas al mismo clúster. En este tipo de clustering considera como irregular a las áreas esparcidas entre clústers. Permite conectar áreas con formas arbitrarias.

Uno de los modelos más usado es el de las K-medias cuyos pasos son:

1. Se eligen k datos aleatorios que serán los centroides representativos de cada clúster.

2. Se analiza la distancia de cada dato al centroide perteneciendo al clúster del centroide más cercano.
3. Se obtiene la media de cada clúster y este será el nuevo centro.
4. Se repiten los pasos 2 y 3 hasta que los clústers no cambien siendo esta una agrupación final a la cual se calcula la varianza total.
5. Se repiten los pasos del 1 al 4 para obtener cuantas agrupaciones finales se desee.
6. Se elige la mejor agrupación a la agrupación final que tenga la menor varianza.

Para saber cual es nuestro número de clústers óptimo (k) usamos el método del codo, el cual consiste en graficar la reducción de la varianza (o bien solo la varianza) vs k y se toma como k al punto en que varianza no disminuirá de forma significativa entre un valor k y otro.

Reglas de asociación

Las reglas de asociación se derivan de un tipo de análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones dentro un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta.

Una regla de asociación se define del tipo "Si A (antecedente) entonces B (consecuencia)": ("$Si\ A \rightarrow B$") donde A y B son ítems individuales.

Las reglas de asociación nos permiten encontrar las combinaciones de artículos que ocurren con mayor frecuencia en una base de datos transaccional, así como medir la fuerza e importancia de estas combinaciones.

Se usan para definir patrones de navegación, promociones de pares de productos, análisis de información de ventas, distribución de mercancías en tiendas, segmentación de clientes con base en patrones de compra, etc.

Existen diferentes tipos de reglas de asociación:

- **Asociación cuantitativa**. Con base en los tipos de valores que manejan las reglas:
  - Asociación Booleana: asociaciones entre la presencia o ausencia de un ítem.
  - Asociación Cuantitativa: describe asociaciones entre ítems cuantitativos o atributos.
- **Asociación multidimensional**. Con base en las dimensiones de datos que involucra una regla:
  - Asociación Unidimensional: los ítems o atributos de la regla se referencian en una sola dimensión.
  - Asociación Multidimensional: los ítems o atributos de la regla se referencian en dos o más dimensiones.

- **Asociación multinivel**. Con base en los niveles de abstracción que involucra la regla:
  - Asociación de un nivel: Los ítems son referenciados en un único nivel de abstracción.
  - Asociación Multinivel: Los ítems son referenciados a varios niveles de abstracción.

Existen algunas métricas o conceptos que hay que tener en claro:

- **Soporte:** número de veces o frecuencia (relativa) con que A y B aparecen juntos en una base de datos de transacciones.

$$Soporte(A \rightarrow B) = P(A \cap B) = \frac{Frecuencia\ en\ que\ A \cap B\ aparece}{Total\ de\ transacciones}$$

  Una regla con bajo soporte puede haber aparecido por casualidad.

- **Confianza:** cociente del soporte de la regla y el soporte del antecedente solamente. Mide la fortaleza de la regla.

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

  En una regla con baja confianza es probable que no exista relación entre antecedente y consecuente.

- **Lift:** refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos sabemos que ocurrió el antecedente.

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

  Un lift mayor a 1 representa relación fuerte y de frecuencia mayor que el azar (complementos). Un lift igual a 1 representa relación del azar. Un lift menor a 1 representa relación débil y de frecuencia menor que el azar (sustitutos).

Detección de outliers

Los outliers o datos atípicos son observaciones que se desvía mucho del resto de las observaciones siendo así una observación sospechosa que se sospecha pudo haber sido generada por mecanismos diferentes al resto de los datos.

Por lo anterior, es fácil ver que los problemas de detección de outliers son problemas de detección de datos raros o de comportamientos inusuales en los datos.

Los outliers pueden significar varias cosas:

- Error.

- Limites. En estos casos hay que mantener el dato para que no perjudique al aprendizaje del modelo.
- Punto de Interés. En caso de que lo que estemos buscando sean casos “anómalos”.

La detección de outliers puede aplicarse a diversas áreas, entre ellas:

- Aseguramiento de ingresos en las telecomunicaciones.
- Detección de fraudes financieros.
- Seguridad y la detección de fallas.

Los outliers pueden afectar a los resultados de un modelo sin embargo eliminarlos no es la solución. Eliminarlos o sustituirlos puede modificar las inferencias que se realicen a partir de esa información, debido a que introduce un sesgo, a que disminuye el tamaño muestral y a que puede afectar tanto a la distribución como a las varianzas. La mejor opción sería explicar esta variabilidad.

Si no es posible explicar la variabilidad de los datos atípicos la mejor opción es quitarles peso a esas observaciones atípicas mediante métodos estadísticos robustos.

Visualización

La visualización de datos es la representación gráfica de información y datos. Es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. Se usan elementos visuales como cuadros, gráficos y mapas.

Según sea la naturaleza de los datos existen diferentes técnicas de aproximación. Podemos clasificar según la complejidad y elaboración de la información de la siguiente manera:

1. **Elementos básicos de representación de datos:** este es el caso más sencillo. Algunos tipos de visualizaciones básicas son:
    a. Gráficas: barras, líneas, columnas, puntos, “tree maps”, tarta, semi-tarta, etc.
    b. Mapas: burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drilldown).
    c. Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.
2. **Cuadros de mando:** son una composición compleja de visualizaciones individuales que tienen coherencia y relación temática entre ellas. Son utilizados para análisis de conjuntos de variables y toma de decisiones.
3. **Infografías:** no están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos, es decir, se utilizan para contar “historias”. Esta

narrativa no se construye a través de texto, sino mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Existen diversos softwares de visualización de datos. Las aplicaciones web son base fundamental para creación de visualizaciones web basadas en datos. Algunos estándares web que se han ido desarrollando en los últimos años para la evolución de las aplicaciones web son HTML5, CSS3, SCV y WebGL.

Predictivas

Este tipo de técnicas nos permiten predecir el valor de un atributo en particular basándonos en los datos recolectados de otros atributos.

Regresión

La regresión predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. Se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática. Puede clasificarse en regresión lineal y no lineal. A su vez, la primera se clasifica en regresión lineal simple y regresión lineal múltiple.

La regresión lineal simple se lleva a cabo cuando sólo se trata de una variable regresora y sigue el modelo:

$$y = \beta_o + \beta_1 x + e$$

Donde e es una variable aleatoria normalmente distribuida con media 0 y varianza σ^2.

Es necesario que la estimación del modelo sea una recta que proporcione un buen ajuste a los datos observados. Para ello se utiliza el ajuste por mínimos cuadrados:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

La regresión lineal múltiple por su parte usa o tiene k regresores, o variables. Su modelo también consiste en una función lineal de parámetros desconocidos:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i = \sum_{i=1}^{k} \beta_j x_{ij} + e_i$$

Así, su función de mínimos cuadrados está dada por:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Al simplificar las derivadas parciales respecto a las β's se obtienen las ecuaciones normales de mínimos cuadrados cuya solución serán los mínimos cuadrados.

Al trabajar con modelos de regresión múltiple en ocasiones es más sencillo manejarlos de forma matricial.

Los usos de las regresiones lineales incluyen, pero no se limitan a: medicina, informática, estadística, comportamiento humano e industria.

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.

Se entrena o estima un modelo usando los datos recolectados para hacer predicciones futuras.

Existen diferentes técnicas de clasificación como lo son:

- **Árboles de decisión:** serie de condiciones organizadas en forma jerárquica, a modo de árbol. Son útiles para problemas que mezclen datos categóricos y numéricos. Se usan en Clasificación, Agrupamiento y Regresión. Se tienen algunos problemas con la inducción de reglas ya que estas podrían no necesariamente formar un árbol, no cubrir todas las posibilidades o entrar en conflicto.
- **Clasificación Bayesiana:** utiliza la regla de Bayes que dice que, si tenemos una hipótesis H sustentada para una evidencia E, entonces $P(H|E) = P(E|H) * \frac{P(H)}{P(E)}$ donde $P(H|E)$ es la probabilidad del suceso H condicionada al suceso E.
- **Redes neuronales:** Trabajan directamente con números (en caso de que se trabaje con datos nominales, estos deben enumerarse). Se usan en Clasificación, Agrupamiento y Regresión. Consisten generalmente de tres capas: de entrada, oculta y de salida. Internamente pueden verse como una gráfica dirigida.
- **Support Vector Machines (SVM)**
- **Clasificación basada en asociaciones**

Patrones secuenciales

Los patrones secuenciales se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias. Es una clase especial de dependencia en donde el orden de acontecimientos es considerado ya que son eventos que se enlazan con el paso del tiempo.

Se trata de buscar asociaciones de la forma “si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n”.

Para analizar los patrones secuenciales se utilizan reglas de asociación secuenciales las expresan patrones de comportamiento que se dan en instantes distintos en el tiempo.

Los patrones secuenciales tienen las siguientes características:

- El orden importa.
- Su objetivo es encontrar patrones en secuencia.
- Una secuencia es una lista ordenada de itemsets, donde cada itemset es un elemento de la secuencia.
- El tamaño de una secuencia es su cantidad de elementos (itemsets).
- La longitud de una secuencia es su cantidad de ítems.
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes (o patrones secuenciales) son las subsecuencias de una secuencia que tienen un soporte mínimo.

Se trabaja usualmente con bases de datos temporales, relacionales o documentales. Sus usos o aplicaciones van desde la medicina hasta finanzas y banca.

Para la resolución de problemas de patrones secuenciales se usa:

- **Agrupación de patrones secuenciales:** separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí, y diferentes a los objetivos de otros grupos. Para esto, primero se selecciona arbitrariamente el centro del primer agrupamiento y luego se procesan secuencialmente los demás patrones mediante cálculos de distancia. Cada M patrones se mezclan agrupamientos, estos pueden ser por cercanía, por tamaño o mezcla forzada.
- **Clasificación con datos secuenciales:** expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos pero cercanos en el tiempo.
- **Reglas de asociación con datos secuenciales:** se dan cuando los datos contiguos presentan algún tipo de relación.

Existen distintos métodos representativos para el análisis de patrones secuenciales como lo son GSP, SPADE, AprioriAll, FreeSpan, SPAM, PrefixSpan, ISM, IncSp, ISE e IncSpan.

Predicción

Para lograr un buen modelo de predicción es necesario tomar en cuenta algunos elementos como lo son la definición del problema, la recopilación de los datos, elección de una medida o indicador de éxito y la preparación de los datos.

En los modelos se dividen los datos de la siguiente manera: 70% para usarlos como conjunto de entrenamiento, un 15% como un conjunto de pruebas y 15% como conjunto de validación.

Algunos de los modelos predictivos más comunes son:

- **Árboles aleatorios o de decisión:** modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Se divide el espacio muestral en subregiones y estas a su vez se pueden dividir en regiones más pequeñas. Están formados por nodos y su lectura se realiza de arriba hacia abajo. En el primer nodo o nodo raíz se produce la primera división en función de la variable más importante. Después se tienen a los nodos internos o intermedios que vuelven a dividir el conjunto de datos en función de las variables y al final, están los nodos terminales u hojas cuya función es indicar la clasificación definitiva. Se clasifican en:

    - **Árboles de regresión:** en los cuales la variable respuesta es cuantitativa. Consiste en hacer preguntas de tipo $¿xk \leq c?$ para cada una de las covariables. Son fáciles de entender e interpretar y requieren poca preparación de los datos. Para asegurarnos que los árboles sean distintos, se usa una estrategia se denomina *bagging* donde cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento.
    - **Árboles de clasificación:** en los cuales la variable respuesta es cualitativa. Consiste en hacer preguntas del tipo $¿xk \leq c?$ para las covariables cuantitativas o preguntas del tipo $¿xk = nivelj?$ para las covariables cualitativas. Hay dos tipos de nodo:
        1. Nodos de decisión: tienen una condición al principio y tienen más nodos debajo de ellos
        2. Nodos de predicción: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo»

        La información de cada nodo es la siguiente:

1. Condición: Si es un nodo donde se toma alguna decisión.
2. Gini: Es una medida de impureza.
   $gini = 1 - \sum_{k=1}^{n}(probabilidad\ de\ cada\ clase)^2$
3. Samples: Número de muestras que satisfacen las condiciones necesarias para llegar a este nodo.
4. Value: Cuántas muestras de cada clase llegan a este nodo.
5. Class: Qué clase se les asigna a las muestras que llegan a este nodo.

- **Bosques aleatorios:** técnica de aprendizaje automático supervisada basada en árboles de decisión. Obtiene un mejor rendimiento de generalización compensando los errores de las predicciones de los distintos árboles de decisión.

Una vez terminado el modelo se debe medir su eficacia para lo que podemos usar **validación cruzada** la cual se emplea para estimar la tasa de error de un modelo y así evaluar su capacidad predictiva.

También existen métricas de eficacia tanto para datos numéricos como categóricos como lo son:

- **Error cuadrático medio:** mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima.
- **Curva ROC:** la cual nos sirve para conocer el rendimiento global de la prueba (área bajo la curva) donde el eje X son los falsos positivos y el eje Y son los verdaderos positivos.