Resumen sobre técnicas de Minería de Datos

Por Mariann Avila

1. *Predicción*

Cuando hablamos de modelos de minería de datos de predicción, en resumen, nos referimos a un proceso en el que se analizan datos actuales históricos reales para obtener información sobre acontecimientos futuros o conocidos. Para cualquiera de los métodos de predicción se necesitan 4 elementos previos: la definición del problema, datos del problema, un indicador de éxito y preparar datos. Además se deben dividir los datos en conjunto de entrenamiento con un 70%, conjunto de validación con 15% y lo restante en el conjunto de pruebas.

Entre las opciones encontramos el "Árbol de decisión" que se caracteriza por dividir en observaciones con valores similares para la variable dependiente, además se forma por nodos y se lee de arriba hacia abajo. Tenemos el "Árbol de regresión" donde se dividen las co variables en hiperrectangulps lo cual hace mas fácil de analizar. Igual existe el "Bosque aleatorio" basada en arboles de decisión, compensa los errores de otros arboles y se puede llamar bagging por usar muestras con reemplazo y luego combinar resultados.

2. *Reglas de Asociación*

Las reglas de asociación se utilizan para encontrar relaciones dentro de un gran conjunto de transacciones entre sus ítems, implica un antecedente y un consecuente. Se puede aplicar en diferentes espacios como segmentar a los clientes dependiendo de sus compras y definir patrones de navegación dentro de la tienda para así realizar acciones como paquetes de promociones o acomodo de artículos.

Existen distintos tipos de reglas de asociación dependiendo de su asociación: Booleana- sobre la presencia o ausencia de un item, Cuantitativa- entre ítems cuantitativos es decir contables, Unidimensional- si los ítems existen en una sola dimensión, Multidimensional- si los ítems se referencian en dos o mas dimensiones, De un Nivel. Un único nivel de atracción o Mulinivel- varios niveles de abstracción

En esta técnica se utilizan tres métricas: Soporte, Confianza y Lift. El soporte nos dice el número de veces o la frecuencia en la que aparecen nuestros. La confianza nos dice la probabilidad de que, si ya escogimos el producto A, escojamos el producto consecuente B. El lift refleja el aumento de probabilidad del consecuente B cuando ya sabemos que escogimos el antecedente A.

3. *Clustering*

La técnica clustering se refiere a agrupar datos y crear particiones según sus similitudes. Se utiliza comúnmente en: Investigación del mercado, Identificar comunidades, Prevención de crimen y Procesamiento de imágenes.

Existen distintos 4 tipos de clustering:

Centroid Based Clustering: Cada cluster es representado por centroides, estos clusters se basan en la distancia de los puntos hasta el centroide, se realizan varias iteraciones hasta llegar al resultado, los centroides son k puntos aleatorios que pasan a ser centroides de cada cluster, de cada dato calculamos su distancia con la del centroide y este dato pertenece al cluster del de la distancia mínima.

Connectivity Based Clustering: Su característica es que un cluster contiene a otros clusters de aquí la llamada jerarquía y que su algoritmo usado es Hierarchical clustering.

Density Based Clustering: Cada cluster pertenece a una distribución normal, los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal y su algoritmo es Gaussian mixture models.

Density Based Clustering: Son definidos por áreas de concentración, trata de conectar puntos que tengan una distancia pequeña, este cluster contiene a todos los puntos relacionados dentro de una distancia limitada

4. *Visualización*

La visualización de datos, aun sonando rebundante, es representar (visualizar) de manera gráfica nuestra información y datos, utilizando elementos visuales como gráficos y mapas, proporcionando una manera más fácil de observar tendencias, valores atípicos y patrones en nuestros datos. Esta técnica es de gran ayuda para analizar cantidades grandes de datos y para la toma de decisiones, es mas fácil observarlos.

En los tipos de visualizaciones se encuentran:

Elementos básicos de representación de datos: es el más sencillo donde se utilizan tipos de visualizaciones básicas como las gráficas, mapas y tablas.

Cuadros de mando: son una composición compleja de visualizaciones individuales que guardan una coherencia y tienen una relación entre ellas

Infografías: aunque no están destinadas para el análisis se utilizan dentro de observación de datos, se utilizan en la construcción de narrativas a partir de los datos, esto quiere decir que se utilizan para contar historias.

Existen softwares de visualización de datos como HTMLS, CSS3, SCV y WebGL, basados en datos generan formas de analizar mejor la situación.

5. *Regresión*

Dentro de la historia de la regresión, en 1805 se registró la primera regresión lineal por el método de los mínimos cuadrados por Legendre y dicho término fue introducido por Francis Galton en su libro "Natural Inheritance" donde analizo la altura de individuos con respecto a su padre abuelos e hijos. Esta técnica tiene, en resumen, el objeto de predecir el valor de un dato basándose en los otros datos que fueron recolectados.

La función de la regresión es encontrar una relación o ecuación matemática mediante el análisis de la variable dependiente (y) y las variables independientes (x's). Existen varios tipos de regresiones, entre ellas se encuentran la regresión lineal simple y la regresión lineal múltiple.

La regresión lineal simple solamente involucra a un regresor (variable dependiente) y una variable independiente, en la cual se tiene como modelo:

$y = B0 + B1x + e$

Mientras que la regresión lineal múltiple involucra k regresores y una variable independiente, en el cual se tiene como modelo:

$y = B0 + B1x1 + B2x2 + B3x3 + \ldots + Bkxk + e$

Ambas regresiones se pueden estimar por la técnica de mínimos cuadrados.

6. *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 (clasifica). Se estima para hacer predicciones futuras.

Existen diferentes técnicas de clasificación para cada objetivo entre las mas conocidas:

La regla de Bayes: que nos dice que $P(H/E) = P(E/H) * P(H)\ P(E)$ en donde P(H) representa la probabilidad del suceso H, P(E) representa la probabilidad del suceso E y P(E/H) representa la probabilidad del suceso E condicionada al suceso H.

Árbol de decisión: serie de condiciones organizadas en forma jerárquica, a modo de árbol

7. *Patrones secuenciales*

La técnica de patrones secuenciales se encarga de analizar los datos y encontrar subsecuencias dentro de un grupo de secuencias, describen el modelo de comprar que un cliente o un grupo de personas realizan relacionando las transacciones efectuadas por ellos en el transcurso del tiempo, dichos eventos se enlazan con el paso del tiempo.

En esta técnica, se buscan asociaciones de que si sucede un evento X en el tiempo t entonces sucederá un evento Y en el tiempo t+n, su objetivo es poder describir las relaciones temporales que existen entre los valores de los atributos del conjunto. En esta técnica el soporte es el porcentaje de secuencias que contienen en un conjunto de secuencias.

La agrupación de patrones secuenciales se encarga de separar en grupos a los datos, en donde los miembros de un grupo sean similares entre sí y sean diferentes a los objetivos de los otros grupos. La clasificación con datos secuenciales expresa patrones de comportamientos secuenciales, en donde se dan en tiempos distintos. La regla de asociación con datos secuenciales presenta la relación que tienen los datos contiguos.

8. *Outliers*

Esta técnica trata sobre la detección de datos raros o comportamientos inusuales en los datos, datos atipicos. Este dato atípico es la observación que se desvía mucho del resto de las observaciones y apareciendo de una manera sospechosa que puede ser generada por mecanismos diferentes al resto de los datos.

Las aplicaciones de esta técnica son principalmente en el aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros y en la seguridad y detección de fallas. Para este método se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos.

La mayoría de los trabajos existentes sobre la detección de datos atípicos yacen en el campo de la estadística. Existen muchas maneras de detectarlos y estos métodos se han diseñado para diferentes circunstancias que son: la distribución de los datos, si los parámetros de distribución son conocidos o no, el número de outliers esperados y el tipo de outliers esperados.