|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| HORA | LUNES | MARTES | MIERCOLES | JUEVES | VIERNES |
| 7 - 9 | DISEÑO Y DESARROLLO  DE ENCUESTAS | MINERÍA DE DATOS | MINERÍA DE  DATOS |  |  |
| 9 - 11 | ESTADÍSTICA MULTIVARIADA |  |  | ESTADÍSTICA  MULTIVARIADA |  |
| 11 –1 |  | META- ANÁLISIS | CONTROL ESTADÍSTICO  DE CALIDAD | DISEÑO Y DESARROLLO  DE ENCUESTAS | CONTROL ESTADÍSTICO  DE CALIDAD |
|  | | | | | |
| 2 - 4 |  |  |  | PROYECTO DE TESIS | META- ANÁLISIS |
| 4 - 6 |  |  |  | INGLES |  |

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos El proceso de descubrimiento de conocimiento en bases de datos 67 de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo predecir para nuevos clientes si son buenos o malos basados en su estado civil, edad, género y profesión, o determinar para nuevos estudiantes si desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas. Entre las tareas predictivas están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los clustering y las correlaciones.

Se han propuesto varios métodos de clasificación: rough sets, árboles de decisión, redes neuronales, Bayes, algoritmos genéticos entre otros. El modelo de clasificación basado en árboles de decisión es probablemente el más utilizado y popular por su simplicidad y facilidad para entender (Han y Kamber, 2001), (Sattler y Dunemann, 2001)

La calidad del árbol depende de la precisión de la clasificación y del tamaño del árbol (Chen, Han y Yu, 1996). El método primero escoge un subconjunto del conjunto de entrenamiento y forma un árbol de decisión. Si el árbol no da la respuesta correcta para todos los objetos del conjunto prueba, una selección El proceso de descubrimiento de conocimiento en bases de datos 69 de excepciones se adiciona al conjunto de entrenamiento y el proceso continúa hasta que se encuentra el conjunto de decisiones correctas. El eventual resultado es un árbol en el cual cada hoja lleva un nombre de la clase y cada nodo interior especifica un atributo con una rama correspondiente a cada posible valor del atributo

• Cada nodo no terminal está etiquetado con un atributo.

• Cada rama que sale de un nodo está etiquetada con un valor de ese atributo.

• Cada nodo terminal está etiquetado con un conjunto de casos, los cuales satisfacen todos los valores de atributos que etiquetan el camino desde ese nodo al nodo inicial.

El algoritmo de clustering segmenta una base de datos sin ninguna indicación por parte del usuario sobre el tipo de clusters que va a encontrar en la base de datos, y desecha cualquier sesgo o intuición por parte del usuario; así potencia el verdadero descubrimiento de conocimiento. Por esta razón, al método de segmentación o clustering se lo denomina aprendizaje no supervisado. Algunos de los algoritmos utilizados para clustering son: K-Means (Han y Kamber, 2001), Clarans (Clustering Large Applications based upon Randomized Search) (Ng y Han, 1994), y Birch (Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang, Ramakrishnan y Livny, 1996).

Asociaci´pn

Una regla de asociación es unidimensional, si los ítems o atributos de la regla hacen referencia a un solo predicado o dimensión. Por ejemplo, se tiene la siguiente regla de asociación:

Los patrones secuenciales, en el dominio de la medicina, se pueden utilizar por ejemplo para ayudar a identificar síntomas y enfermedades que preceden a otras enfermedades.

La metodología crisp-dm consiste en un conjunto de tareas que están organizadas en cuatro niveles de abstracción: fases, tareas generales, tareas especializadas e instancias de proceso (véase figura 2). Dichos niveles están establecidos respetando jerarquías en tareas; inician en el nivel más general hasta llegar, finalmente, a los casos más específicos (Chapman et al., 2000).

Por ejemplo, si el problema es de clasificación, podemos elegir entre árboles de decisión, k-nearest neighbour o razonamiento basado en caos (cbr).

• Generar plan de prueba. Se debe generar un plan para probar la calidad y validez del modelo construido; por ejemplo, en una tarea como la clasificación es posible usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba.

• Construir el modelo. Se ejecuta la técnica seleccionada sobre los datos preparados para generar uno o más modelos. Todas las técnicas del modelado tienen un conjunto de parámetros que determinan características del modelo