Transformación y Reducción de datos

Pasos en pre-processing

Se deben aplicar algunos pasos de preprocesamiento para que los datos sean más adecuados para el análisis de datos.

- Agregación
- Muestreo
- Reducción de dimensionalidad
- Selección de subconjunto de características
- Creación de características
- Discretización y binarización
- Transformación de variables

Agregación

Ejemplos:

- Los atributos cuantitativos, como el precio, generalmente se agregan tomando una suma o un promedio.

- Un atributo cualitativo, como un artículo, puede omitirse o resumirse como el conjunto de todos los artículos que se vendieron en esa ubicación.

- Reduciendo los valores posibles para la fecha de 365 días a 12 meses.

Muestreo

- Una muestra es representativa si tiene aproximadamente la misma propiedad (de interés) que el conjunto original de datos.

- Si la media (promedio) de los objetos de datos es la propiedad de interés, entonces una muestra es representativa si tiene una media cercana a la de los datos originales.

Muestreo

1. Muestreo aleatorio simple
    1. muestreo sin reemplazo
    2. muestreo con reemplazo
2. Muestreo estratificado :

    se extrae la misma cantidad de objetos de cada grupo aunque los grupos sean de diferentes tamaños.
3. Muestreo adaptativo o progresivo:

    comience con una muestra pequeña y luego aumente el tamaño de la muestra hasta que se haya obtenido una muestra de tamaño suficiente.

Reducción de dimensionalidad

- El término *reducción de dimensionalidad* a menudo se reserva para aquellas técnicas que reducen la dimensionalidad de un conjunto de datos mediante la creación de *nuevos atributos* que son *una combinación de los atributos antiguos*.

The Curse of Dimensionality: se refiere al fenómeno de que muchos tipos de análisis de datos se vuelven significativamente más difíciles a medida que aumenta la dimensionalidad de los datos.

Reducción de dimensionalidad

Principal Components Analysis (PCA)

una técnica de álgebra lineal para atributos continuos que encuentra nuevos atributos (componentes principales) que:

1. son combinaciones lineales de los atributos originales
2. son ortogonales (perpendiculares) entre sí
3. capturan la cantidad máxima de variación en los datos

Por ejemplo, los primeros dos componentes principales capturan tanta variación en los datos como sea posible con dos atributos ortogonales que son combinaciones lineales de los atributos originales.

Reducción de dimensionalidad

Singular Value Decomposition (SVD):

Descomposición de valores singulares (SVD) es una técnica de álgebra lineal que está relacionada con PCA y también se usa comúnmente para la reducción de dimensionalidad.

Selección de subconjunto de características

Características redundantes:

Ej: El precio de compra de un producto y el monto del impuesto sobre las ventas pagado contienen gran parte de la misma información.

Funciones irrelevantes:

Ej: Los números de identificación de los estudiantes son irrelevantes para la tarea de predecir los promedios de calificaciones de los estudiantes.

Las características redundantes e irrelevantes pueden reducir la precisión de la clasificación y la calidad de los grupos que se encuentran.

Selección de subconjunto de características

1. Enfoques integrados :

La selección de características ocurre naturalmente como parte del algoritmo de analisis de datos.

Específicamente, durante la operación del algoritmo de analisis de datos, el propio algoritmo decide qué atributos usar y cuáles ignorar.

Selección de subconjunto de características

2. Enfoques de filtrado

Las funciones se seleccionan antes de que se ejecute el algoritmo de analisis de datos, utilizando algún enfoque que sea independiente de la tarea de analisis de datos.

Ej: Podríamos seleccionar conjuntos de atributos cuya correlación por pares sea lo más baja posible.

Selección de subconjunto de características

3. Enfoques de rapero envolvente

Como una caja negra para encontrar el mejor subconjunto de atributos,

pero normalmente sin enumerar todos los subconjuntos posibles.

Selección de subconjunto de características

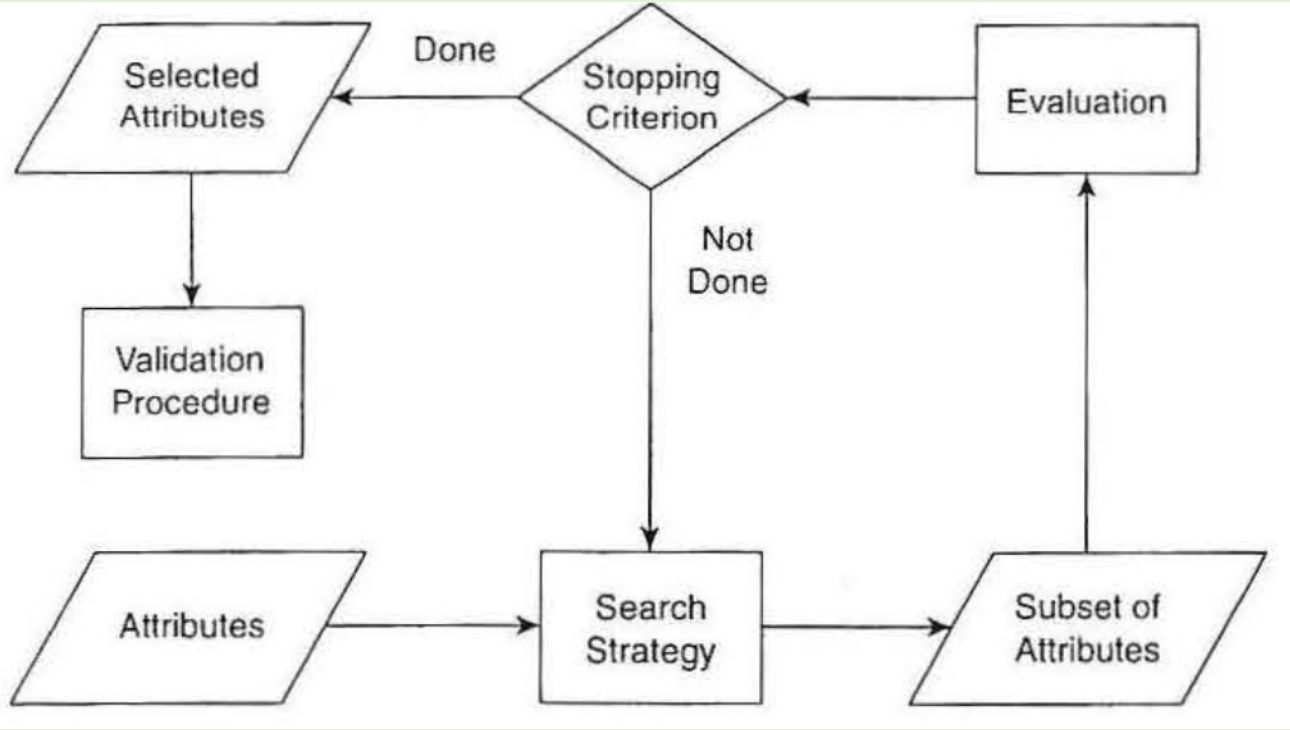

Diagrama de flujo de un proceso de selección de un subconjunto de características.

Creación de características

Metodologías para la creación de nuevos atributos:

- extracción de características
- mapear los datos a un nuevo espacio
- construcción de características

Creación de características - extracción de características

La creación de un nuevo conjunto de características a partir de los datos sin procesar originales se conoce como *extracción de características*.

Ej: Considere un conjunto de fotografías, donde cada fotografía debe ser clasificada de acuerdo a si contiene o no a. cara humana. Los datos sin procesar son un conjunto de píxeles y, como tales, no son adecuados para muchos tipos de algoritmos de clasificación. Sin embargo, si los datos se procesan para proporcionar características de mayor nivel, como la presencia o ausencia de ciertos tipos de bordes y áreas que están altamente correlacionados con la presencia de rostros humanos, entonces se puede aplicar un conjunto mucho más amplio de técnicas de clasificación para este problema.

Creación de características - mapear los datos a un nuevo espacio

Una visión totalmente diferente de los datos puede revelar características importantes e interesantes.

Ej: Considere los datos de series de tiempo, que a menudo contienen patrones periódicos.

- Si solo hay un único patrón periódico y no hay mucho ruido, entonces el patrón se detecta fácilmente.
- Si, por otro lado, hay una serie de patrones periódicos y hay una cantidad significativa de ruido presente, entonces estos patrones son difíciles de detectar. Sin embargo, dichos patrones pueden detectarse a menudo aplicando una transformada de Fourier a la serie temporal para cambiar a una representación en la que la información de frecuencia sea explícita.

Creación de características - construcción de características

A veces, las características de los conjuntos de datos originales tienen la información necesaria, pero no está en una forma adecuada para el algoritmo de analisis de datos.

En esta situación, una o más funciones nuevas construidas a partir de las funciones originales pueden ser más útiles que las funciones originales.

Habido algunos intentos de realizar automáticamente la construcción de características mediante la exploración de combinaciones matemáticas simples de atributos existentes (densidad = masa/volumen), el enfoque más común es construir características utilizando la experiencia del dominio (Domain knowledge).

Discretización y binarización

Algunos algoritmos de análisis de datos, especialmente ciertos algoritmos de clasificación, requieren que los datos estén en forma de atributos categóricos.

Los algoritmos que encuentran patrones de asociación requieren que los datos estén en forma de atributos binarios.

Por lo tanto,

- a menudo es necesario transformar un atributo continuo en un atributo categórico (discretización)
- es posible que sea necesario transformar tanto los atributos continuos como los discretos en uno o más atributos binarios (binarización)

Transformación de variables

Se refiere a una transformación que se aplica a todos los valores de una variable.

- Si x es una variable, los ejemplos de tales transformaciones incluyen

  x^k, log x, e^x, sqrt (X), 1/x, sen(x) o lxl.

Transformación de variables

Normalización o Estandarización

El objetivo es hacer que un conjunto completo de valores tenga una propiedad particular.

Si $\underline{x}$ es la media (promedio) de los valores de atributo y *Sx* es su desviación estándar de *poblacion*, entonces la transformación x' = $(x - \underline{x})/Sx$ crea una nueva variable que tiene una media de 0 y una desviación estándar de 1 .