CURSO ESPECIALIZACIÓN IA Y BIG DATA

MÓDULO: SISTEMAS DE APRENDIZAJE AUTOMÁTICO

U2: PREPARACIÓN DE DATOS

L'FSE inverteix en el teu futur

Fons Social Europeu - FSE

UNIÓ EUROPEA

GENERALITAT VALENCIANA

Conselleria d'Educació, Cultura i Esport

Índice

3.3.- Selección de datos

Casi nunca es posible estudiar a toda la población que te interesa, es por esto que se utilizan diversos tipos de muestreo a la hora de recopilar datos.

El proceso de muestreo y selección de datos es esencial para garantizar que los datos utilizados en proyectos de análisis de datos sean precisos y representativos.

- Identificar los datos a utilizar garantizando que los datos sean representativos y precisos.
- Determinar el tamaño de la muestra.
- Selección de la muestra garantizando que la muestra sea representativa y los datos no estén sesgados.
- Validar la muestra.

3.3.- Selección de datos

Tipos de muestreo horizontal:

- Muestreo Aleatorio Simple
- Muestreo Aleatorio Estratificado
- Muestreo de Grupos
- Muestreo Exhaustivo

3.3.- Selección de datos

Muestreo Aleatorio Simple:

Cualquier instancia tiene la misma probabilidad de ser extraída en la muestra.

Muestreo Aleatorio Estratificado:

El objetivo de este muestreo es obtener una muestra balanceada con suficientes elementos de todos los estratos, o clases.
Una versión simple es realizar un muestreo aleatorio simple sin reemplazamiento de cada estrato hasta obtener los n elementos de ese estrato.
Si no hay suficientes elementos en un estrato podemos utilizar en estos casos muestreo aleatorio simple con reemplazamiento (sobremuestreo).

3.3.- Selección de datos

Muestreo de Grupos:
El muestreo de grupos consiste en elegir sólo elementos de unos grupos.
Objetivo: Descartar ciertos grupos que, por diversas razones, pueden impedir la obtención de buenos modelos.

Muestreo Exhaustivo:
Para los atributos numéricos (normalizados) se genera al azar un valor en el intervalo posible y se elige el más cercano.
Para los atributos nominales se genera al azar un valor entre los posibles.
Con esto obtenemos una instancia ficticia y buscamos la instancia real más similar a la ficticia. Se repite este proceso hasta tener n instancias.
Objetivo: Cubrir completamente el espacio de instancias.

Podemos hacer combinaciones con las técnicas de muestreo

3.3.- Selección de datos

¿Con cuantos datos es preferible trabajar?

Depende de:

- El número de atributos y valores que tomen estos (“grados de libertad”)
- El método de aprendizaje y de su expresividad (por ejemplo una regresión lineal requiere muchos menos ejemplos que una red neuronal)

Se utiliza una estrategia incremental, en el que se va haciendo la muestra cada vez más grande (y diferente si es posible) hasta que se vea que los resultados no varían significativamente entre un modelo y otro.

3.3.- Selección de características

Objetivos:

- **Muchas técnicas de minería de datos pueden perderse con muchos atributos, especialmente si estos atributos no disponen de un número suficiente de ejemplos**
- Reducir el tamaño de los datos
- Hacer una buena selección puede mejorar la calidad del modelo
- Permite expresar el modelo resultante en función de menos variables
- Representar mejor los datos visualmente
- Muchos datos erróneos o faltantes

Reglas generales:

- Eliminación de cualquier atributo que pueda ser clave candidata de una tabla.
- Eliminación de atributos dependientes

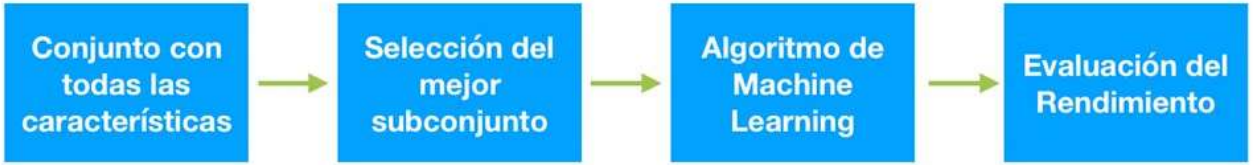
3.3.- Selección de características

Existen dos tipos generales de métodos para seleccionar características:

Métodos de filtro o métodos previos: se filtran los atributos irrelevantes antes de cualquier proceso de minería de datos

Métodos basados en modelo o métodos de envolvente (*wrapper*): la bondad de la selección de atributos se evalúa respecto a la calidad de un modelo de extraído a partir de los datos (utilizando, lógicamente, algún buen método de validación).

Métodos de filtro



↓Características/Predicción →	Continuo	Categórico
Continuo	Correlación de Pearson	LDA
Categórico	Anova	Chi-cuadrado

Métodos de filtro

Correlación de Pearson: se usa como una medida para cuantificar la dependencia lineal entre dos variables continuas X e Y, su valor varía de -1 a +1.

LDA: el análisis discriminante lineal se usa para encontrar una combinación lineal de características que caracteriza o separa dos o más clases, o niveles, de una variable categórica.

ANOVA: significa análisis de la varianza y es similar a LDA, excepto por el hecho de que opera mediante una o más funciones independientes categóricas y una función dependiente continua. Proporciona una prueba estadística de si las medias de varios grupos son iguales o no.

Chi-cuadrado: es una prueba estadística que se aplica a los grupos de características categóricas para evaluar la probabilidad de correlación o asociación entre ellos utilizando su distribución de frecuencia.

Métodos de envolvente

Métodos de envolvente. Ejemplos

Forward Selection: es un método iterativo en el que comenzamos sin tener ninguna característica en el modelo. En cada iteración, seguimos agregando la función que mejor mejora nuestro modelo hasta que la adición de una nueva variable no mejore el rendimiento del modelo.

Backward Selection: comenzamos con todas las características y eliminamos la característica menos significativa en cada iteración, lo que mejora el rendimiento del modelo. Repetimos esto hasta que no se observe ninguna mejora en la eliminación de características.

Recursive Feature Elimination: es un algoritmo de optimización que busca encontrar el subconjunto de funciones con mejor rendimiento. Crea repetidamente modelos y deja de lado la mejor o la peor característica de rendimiento en cada iteración. Construye el siguiente modelo con las características de la izquierda hasta que se agotan todas las características, luego clasifica las características según el orden de su eliminación.

Métodos integrados

Combina las cualidades de los métodos de filtro y envoltura. Se implementa mediante algoritmos que tienes sus propios métodos de selección de características incorporados.

Algunos de los ejemplos más populares de estos métodos son la regresión LASSO y RIDGE, que tienen funciones de penalización incorporadas para reducir el sobreajuste.

Gracias