Preprocesado de Datos

Juan A. Botía

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

Ingeniería Superior en Informática, UMU

Data Preprocessing

Vamos a intentar responder a la pregunta

Preprocesado de datos

¿cómo puedo preparar los datos para que el proceso de minería de datos (exploratorio o no) sea más fácil y efectivo?

Tareas en el preproceso de datos

Nos vamos a encontrar varios tareas propias de preprocesado de datos:

- Data cleaning o limpieza de datos, es el proceso orientado a eliminar datos con ruido o incorrectos,
- Data integration, trata de integrar diferentes fuentes de datos en un almacén coherente y homogéneo como un *data warehouse* o un *data cube*.
- Data transformation, o transformación de los datos como, por ejemplo, una normalización.
- Data reduction, o reducción de los datos, orientado a reducir el tamaño de los datos mediante la agregación y/o eliminación de características redundantes, o clustering.

Data Cleanning

Los datos del mundo real suelen presentarse en una forma incompleta, inconsistente y ruidosa, limpiaremos los datos mediante

- eliminando datos que faltan,
- suavizando el efecto del ruido,
- eliminando datos fuera de rango y
- corrigiendo inconsistencias.

Qué hacer ante los datos nulos

Recuerda que antes de ponernos manos a la obra con los datos nulos

> *Hay que analizar la razón de su existencia!! ¿Es significativo el hecho de que aparezcan nulos en los datos?*

Diferentes tratamientos de valores nulos

Las siguientes son formas de enfrentarnos a los valores nulos

- Ignorar la tupla (i.e. el dato que falta es el de la clase o existen nulos en varios de los atributos)
- Rellenar los datos de manera manual (inabordable en el data mining)
- Usar una constante global para rellenar los datos (i.e. cambiar cada ausente por `unknown`)
- Usar la media del atributo para sustitución
- Usar la media del atributo obtenida con todos los ejemplares que pertenecen a la misma clase que el ejemplar a modificar
- Utilizar técnicas de minería de datos para predecir el valor más probable en cada caso (e.g. un árbol de decisión)

Comentarios a los métodos

- Todos los métodos con sustitución automática introducen un sesgo en los datos (i.e. la sustitución puede no ser correcta)
- Usar minería es uno muy popular
  - Se aprovecha la información que se tiene de los datos presentes para predecir valores ausentes
  - Aumentamos la posibilidad de que el patrón sea el mismo que se obtiene con todos los datos [1]
- Para profundizar, el libro de texto de Pyle [2], viene con un estupendo material en el capítulo 8, dedicado a esta problemática particular.

Datos con ruido

El ruido en los datos se ve, tradicionalmente, como un error en las variables que se están midiendo.

Este error se modela como una v.a., dada la v.a. X, se aproxima el valor de dicha variable con la expresión

$$\hat{X} = X + e$$

¿Cómo podemos suavizar los valores de $\hat{X}$ de tal forma que minimicemos el error e? Tendremos binning, clustering y regresión

Métodos para aliviar el ruido

El *binning*.

En qué consiste en binning

Una serie de valores ordenados se agrupan en porciones y luego *se suaviza* cada porción. De esta forma, lo que se hace es un tratamiento local del ruido ya que se actúa de manera individual en cada porción.

Un ejemplo

$$4, 8, 15, 21, 21, 24, 25, 28, 34$$

Los dividimos en tres bins de la misma longitud

Bin 1: 4, 8, 15, Bin 2: 21, 21, 24, Bin 3: 25, 28, 34.

El binning

Si sustituimos cada valor por la media, estamos suavizando mediante medias y nos queda

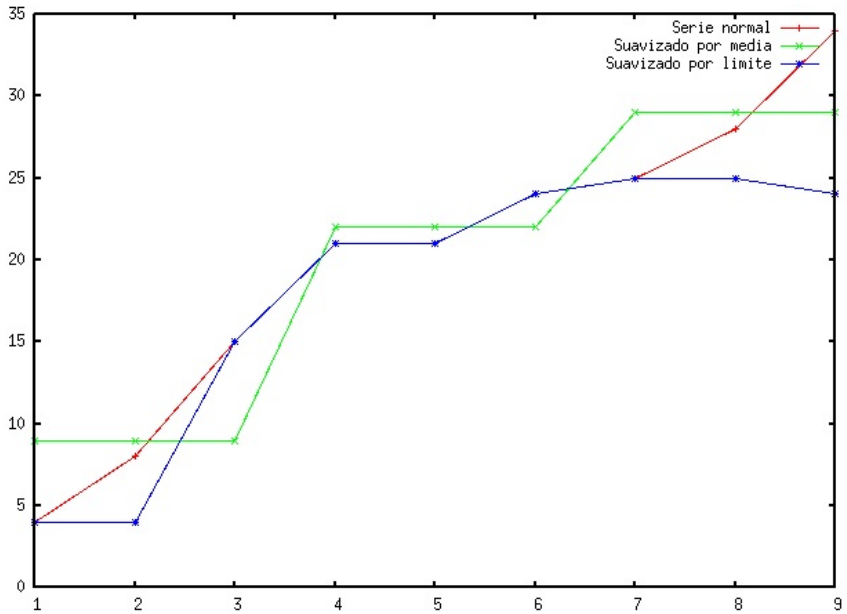
Bin 1: 9, 9, 9, Bin 2: 22, 22, 22, Bin 3: 29, 29, 29.

Ahora, si sustituimos los valores de cada porción por su valor del extremo del bin más cercano, tenemos

Bin 1: 4, 4, 15, Bin 2: 21, 21, 24, Bin 3: 25, 25, 34.

El binning (y II)

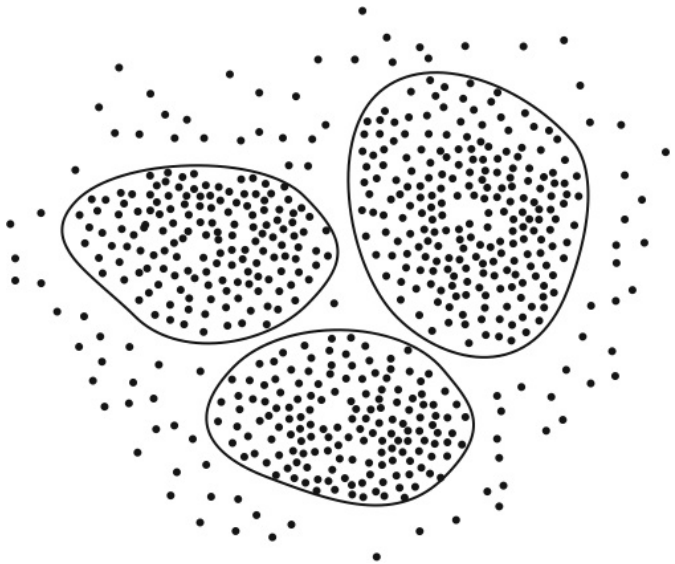
Si ahora los representamos vemos que efectivamente, las curvas se suavizan



- El efecto del binning está en relación a la longitud de las porciones
- Podemos hacer división de la misma anchura (i.e. el mismo rango de valores por cada intervalo).

Métodos para aliviar el ruido (y II)

Podemos reducir el ruido mediante procesos de clustering.



Aquellos valores que caen fuera de los clusters pueden considerarse como valores fuera de rango

Métodos para aliviar el ruido (y III)

Podeos reducir el ruido con la regresión lineal

- Dado un conjunto de tuplas representado por dos variables, hallamos la línea recta que mejor se ajusta a esos datos
- Minimizamos un error basado en el cuadrado de diferencias cuadráticas entre puntos de la recta y reales
- Si el número de variables es mayor que dos tenemos **regresión lineal múltiple**
- En realidad buscamos una expresión matemática con la que reproducir los datos y elimar, así, el error en la medida de lo posible

Data Transformation

Tranformamos unos datos en otros equivalentes y así dejarlos listos para la minería

- Suavizado o eliminación del ruido (utilizamos binning, clustering, regresión, etc)
- Agregación: agregamos valores de atributos, e.g. agregamos ventas diarias en semanales y/o mensuales (i.e. ver los data cubes, en donde se usa típicamente este tipo de transformación).
- Generalización: mediante jerarquías de conceptos, sustituimos valores categóricos o numéricos por otros valores más abstractos (e.g. calle por ciudad, 30 por mediana edad).
- Normalización: escalamos el atributo a un cjto. de valores apropiado según el caso
- Construcción de atributos: mediante el cual construimos nuevos atributos cuando esto es conveniente para el proceso de minería.

Normalización

Las formas más comunes de normalización

- Tenemos la normalización min/max, mediante la expresión

$$v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A.$$

- Normalización de media cero (o z normalisation)

$$v' = \frac{v - \overline{A}}{\sigma_A},$$

en donde $\overline{A}$ y σ_A son la media y desviación estándar, respectivamente.

- Normalización mediante escalado decimal
  - Se mueve el punto decimal en los valores del atributo
  - Las posiciones dependen del máximo en valor absoluto tal que

$$v' = \frac{v}{10^j},$$

en donde j es el entero más pequeño tal que $Max(|v'|) < 1$

Construcción de atributos

Atributos nuevos, ¿para qué y cómo?

- El añadir atributos ayuda a poner a los algoritmos la tarea de análisis un poco más fácil
- Si podemos combinar atributos mediante alguna expresión interesante, conseguimos que el algoritmo no tenga que descubrir dicha expresión
- Ejemplo: podemos crear un atributo área a partir de la altura y la anchura

Construcción de atributos

¿Para qué

Cuando el cjto. de datos es realmente grande, es posible que las técnicas de minería de datos disponibles no sean adecuadas, hasta tal punto que sea totalmente imposible realizar un proceso de minería

¿Cómo?

Las siguientes son formas de reducir los datos que podemos aplicar:

- Agregación en data cubes
- Reducción de dimensiones
- Compresión de datos
- Reducción de la numerosidad (i.e. del número de tuplas)
- Discretización de atributos y generación de jerarquías de conceptos.

Libros aconsejables


Jiawei Han and Micheline Kamber.
Data Mining. Concepts and Techniques.
Morgan Kauffman Publisher, 2001.


Dorian Pyle.
Data Preparation for Data Mining.
Morgan Kauffman Publisher, 1999.