Tipo y Calidad de los Datos

Minería de Datos e Inteligencia de Negocios

Juan Luis Restituyo

Introducción 2

Tipos de Datos 2

- Atributos y medidas 3
- Diferentes tipos de atributos 3
- Tipos de conjuntos de datos 4
- Características generales de los conjuntos de datos 4
- Registros 5
- Datos basados en gráficos 7

Calidad de los datos 11

- Errores de medición y recopilación de datos 11
- Ruido y artefactos 11
- Precisión, sesgo y exactitud 12
- Valores atípicos: 12
- Valores faltantes 13
- Eliminar objetos de datos o atributos 13
- Estimar valores faltantes 13
- Ignorar el valor faltante durante el análisis 13
- Valores Inconsistentes 14
- Datos duplicados 14
- Problemas relacionados con las aplicaciones 14
- Conocimiento sobre los datos. 15

Introducción

En esta lectura vamos a tratar varios problemas relacionados con los datos que son importantes para la minería de datos exitosa:

El tipo de datos: Los conjuntos de datos difieren de varias maneras. Por ejemplo, los atributos utilizados para describir objetos de datos pueden ser de diferentes tipos cuantitativos o cualitativos, y los conjuntos de datos pueden tener características especiales; Por ejemplo, algunos conjuntos de datos contienen series de tiempo u objetos con relaciones explícitas entre sí. No es sorprendente que el tipo de datos determine qué herramientas y técnicas pueden usarse para analizar los datos. Además, las nuevas investigaciones en minería de datos a menudo se deben a la necesidad de acomodar nuevas áreas de aplicación y sus nuevos tipos de datos.

La calidad de los datos: los datos a menudo están lejos de ser perfectos. Si bien la mayoría de las técnicas de extracción de datos pueden tolerar algún nivel de imperfección en los datos, un enfoque en la comprensión y mejora de la calidad de los datos generalmente mejora la calidad del análisis resultante. Los problemas de calidad de los datos que a menudo deben abordarse incluyen la presencia de ruido y valores atípicos; datos faltantes, inconsistentes o duplicados; y datos sesgados o, de alguna otra manera, no representativos del fenómeno o población que se supone que los datos describen.

Tipos de Datos

Un conjunto de datos a menudo se puede ver como una colección de objetos de datos. Otros nombres para un objeto de datos son registro, punto, detector, patrón, evento, caso, muestra, observación o entidad. A su vez, los objetos de datos se describen mediante una serie de atributos que capturan las características básicas de un objeto, como la masa de un objeto físico o el momento en que ocurrió un evento. Otros nombres para un atributo son variable, característica, campo o dimensión.

Ejemplo: (Información del estudiante). A menudo, un conjunto de datos es un archivo, en el que los objetos son registros (o filas) en el archivo y cada campo (o columna) corresponde a un atributo. Por ejemplo, la siguiente tabla muestra un conjunto de datos que consta de información del estudiante. Cada fila corresponde a un estudiante y cada columna es un atributo que describe algún aspecto de un estudiante, como el promedio de calificaciones (GPA) o el número de identificación (ID).

Table 2.1. A sample data set containing student information.

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

Atributos y medidas

En esta sección, abordamos el tema de la descripción de datos considerando qué tipos de atributos se utilizan para describir los objetos de datos. Primero definimos un atributo, luego consideramos qué entendemos por tipo de atributo y finalmente describimos los tipos de atributos que se encuentran comúnmente.

Un atributo es una propiedad o característica de un objeto que puede variar, ya sea de un objeto a otro o de un momento a otro.

Diferentes tipos de atributos

Los atributos nominales y ordinales se denominan colectivamente como atributos categóricos o cualitativos. Como su nombre lo indica, los atributos cualitativos, como la identificación del empleado, carecen de la mayoría de las propiedades de los números. Incluso si están representados por números.

Los atributos cuantitativos están representados por números y tienen la mayoría de las propiedades de los números. Tenga en cuenta que los atributos cuantitativos pueden ser de valores enteros o continuos.

Una forma independiente de distinguir entre atributos es por el número de valores que pueden tomar.

Discreto: un atributo discreto tiene un conjunto de valores finitos o infinitamente contables. Dichos atributos pueden ser categóricos, como códigos postales o números de identificación, o numéricos, como conteos. Los atributos discretos a menudo se representan utilizando variables enteras. Los atributos binarios son un caso especial de atributos discretos y suponen solo dos valores, por ejemplo, verdadero / falso, sí / no, masculino / femenino o 0/1. Los atributos binarios a menudo se representan como variables booleanas o como variables enteras que solo toman el valor valores 0 o 1.

Continuo: un atributo continuo es uno cuyos valores son números reales. Los ejemplos incluyen atributos tales como temperatura, altura o peso. Los atributos continuos se representan típicamente como variables de punto flotante. En la práctica, los valores reales solo pueden medirse y representarse con precisión limitada.

Tipos de conjuntos de datos

Hay muchos tipos de conjuntos de datos, y a medida que el campo de la minería de datos se desarrolla y madura, una mayor variedad de conjuntos de datos estará disponible para el análisis. En esta sección, describimos algunos de los tipos más comunes. Para mayor comodidad, hemos agrupado los tipos de conjuntos de datos en tres grupos: datos de registro, datos basados en gráficos y datos ordenados. Estas categorías no cubren todas las posibilidades y otras agrupaciones son ciertamente posibles.

Características generales de los conjuntos de datos

Antes de proporcionar detalles de tipos específicos de conjuntos de datos, analizamos tres características que se aplican a muchos conjuntos de datos y tienen un impacto significativo en las técnicas de extracción de datos que se utilizan: dimensionalidad, dispersión y resolución.

Dimensionalidad: la dimensionalidad de un conjunto de datos es el número de atributos que poseen los objetos en el conjunto de datos. Los datos con un pequeño número de dimensiones tienden a ser cualitativamente diferentes a los datos de dimensiones moderadas o altas. De hecho, las dificultades asociadas con el análisis de datos de alta dimensión a veces se conocen como la maldición de la dimensionalidad. Debido a esto, una motivación importante en el preprocesamiento de los datos es la reducción de la dimensionalidad.

Disperción: Para algunos conjuntos de datos, como aquellos con características asimétricas, la mayoría de los atributos de un objeto tienen valores de 0; en muchos casos, menos del 1% de las entradas no son cero. En términos prácticos, la dispersión es una ventaja porque generalmente solo los valores que no son cero necesitan ser almacenados y manipulados. Esto se traduce en ahorros significativos con respecto al tiempo de cálculo y almacenamiento. Además, algunos algoritmos de minería de datos funcionan bien solo para datos dispersos.

Resolución: con frecuencia es posible obtener datos a diferentes niveles de resolución y, a menudo, las propiedades de los datos son diferentes en diferentes resoluciones. Por ejemplo, la superficie de la Tierra parece muy desigual a una resolución de unos pocos metros, pero es relativamente suave a una resolución de decenas de kilómetros. Los patrones en los datos también dependen del nivel de resolución. Si la resolución es demasiado fina, un patrón puede no ser visible o puede estar enterrado en el ruido; Si la resolución es demasiado aproximada, el patrón puede desaparecer. Por ejemplo, las variaciones en la presión atmosférica en una escala de horas reflejan el movimiento de las tormentas y otros sistemas climáticos. En una escala de meses, tales fenómenos no son detectables.

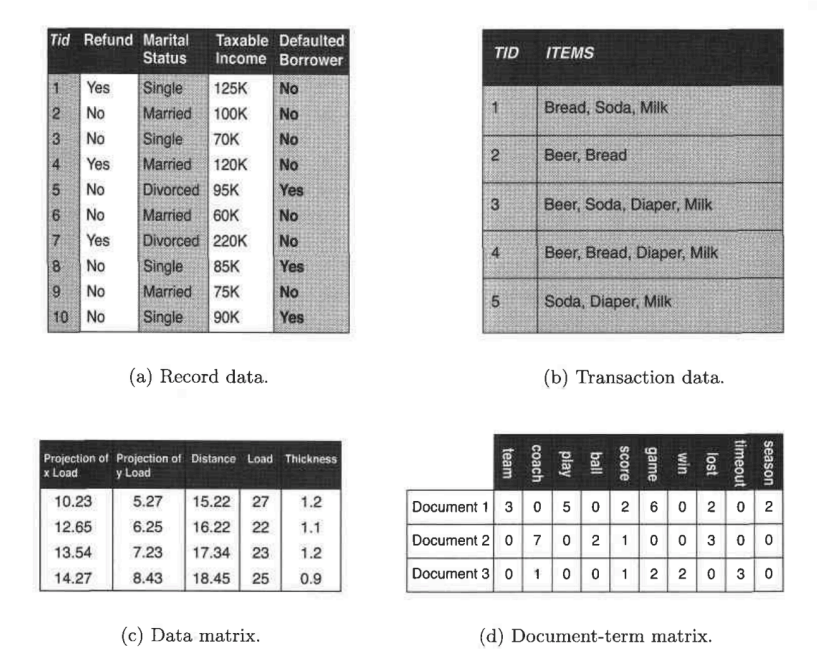
Registros

Gran parte del trabajo de minería de datos supone que el conjunto de datos es una colección de registros (objetos de datos), cada uno de los cuales consta de un conjunto fijo de campos de datos (atributos). Para la forma más básica de datos de registro, no existe una relación explícita entre registros o campos de datos, y cada registro (objeto) tiene el mismo conjunto de atributos. Los datos de registro generalmente se almacenan en archivos planos o en bases de datos relacionales. Las bases de datos relacionales son ciertamente más que una colección de registros, pero la extracción de datos a menudo no utiliza ninguna de la información adicional disponible en una base de datos relacional. Más bien, la base de datos sirve como un lugar conveniente para encontrar registros.

Transacción o datos de la cesta de mercado: Los datos de transacción son un tipo especial de datos de registro, donde cada registro (transacción) involucra un conjunto de elementos. Considere la posibilidad de una tienda de comestibles. El conjunto de productos comprados por un cliente durante un viaje de compras constituye una transacción, mientras que los productos individuales que se compraron son los artículos. Este tipo de datos se denomina datos de la cesta de mercado porque los elementos de cada registro son los productos de la "cesta de mercado" de una persona. Los datos de transacción son una colección de conjuntos de elementos, pero se pueden ver como un conjunto de registros cuyos campos son atributos asimétricos. La mayoría de las veces, los atributos son binarios, que indican si un artículo se compró o no, pero, en términos más generales, los atributos pueden ser discretos o continuos, como la cantidad de artículos comprados o la cantidad gastada en esos artículos. Cada fila representa las compras de un cliente en particular en un momento determinado.

La matriz de datos: si los objetos de datos en una colección de datos tienen todos el mismo conjunto fijo de atributos numéricos, entonces los objetos de datos se pueden considerar como puntos (vectores) en un espacio multidimensional, donde cada dimensión representa un atributo distinto que describe el objeto. Un conjunto de tales objetos de datos puede interpretarse como una matriz m por n, donde hay m filas, una para cada objeto, y n columnas, una para cada atributo. (Una representación que tiene objetos de datos como columnas y atributos como filas también está bien). Esta matriz se denomina matriz de datos o matriz de patrones. Una matriz de datos es una variación de los datos de registro, pero como se trata de atributos numéricos, la operación de matriz estándar se puede aplicar para transformar y manipular los datos. Por lo tanto, la matriz de datos es el formato de datos estándar para la mayoría de los datos estadísticos.

La matriz de datos dispersos: una matriz de datos dispersos es un caso especial de una matriz de datos en la que los atributos son del mismo tipo y son asimétricos; es decir, solo los valores distintos de cero son importantes. Los datos de transacción son un ejemplo de una matriz de datos dispersos que solo tiene 0 1 entradas. Otro ejemplo común es la información del documento. En particular, si se ignora el orden de los términos (palabras) en un documento, entonces un documento se puede representar como un vector de término, donde cada término es un componente (atributo) del vector y el valor de cada componente es el número Muchas veces el término correspondiente aparece en el documento. Esta representación de una colección de documentos a menudo se denomina matriz de término de documento. Los documentos son las filas de esta matriz, mientras que los términos son las columnas. En la práctica, solo se almacenan las entradas que no son cero de las matrices de datos dispersos.



Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	*ITEMS*
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

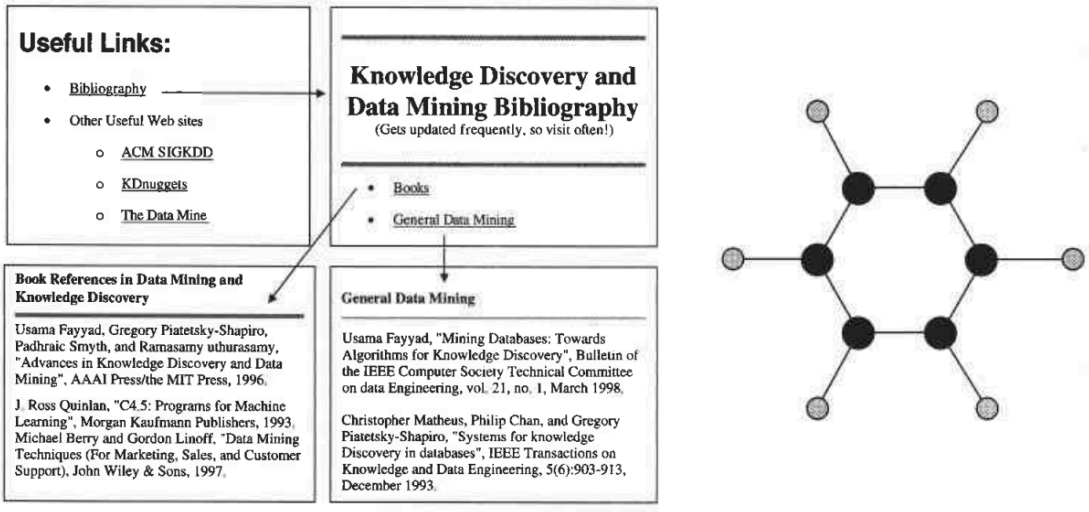
Figure 2.2. Different variations of record data.

Datos basados en gráficos

Un gráfico a veces puede ser una representación conveniente y poderosa para datos, consideramos dos casos específicos:
- El gráfico captura las relaciones entre los objetos de datos
- Los objetos de datos en sí se representan como gráficos.

Datos con relaciones entre objetos: las relaciones entre objetos frecuentemente transmiten información importante. En tales casos, los datos se representan a menudo como un gráfico. En particular, los objetos de datos se asignan a los nodos del gráfico, mientras que las relaciones entre los objetos se capturan mediante los enlaces entre los objetos y las propiedades de los enlaces, como la dirección y el peso. Considere las páginas web en la World Wide Web, que contienen tanto texto como enlaces a otras páginas. Para procesar las consultas de búsqueda, los motores de búsqueda web recopilan y procesan páginas web para extraer sus contenidos. Sin embargo, es bien sabido que los enlaces hacia y desde cada página brindan una gran cantidad de información sobre la relevancia de una página web para una consulta y, por lo tanto, también se debe tener en cuenta.

Datos con objetos que son gráficos: si los objetos tienen una estructura, es decir, los objetos contienen sub-objetos que tienen relaciones, entonces dichos objetos se representan con frecuencia como gráficos. Por ejemplo, la estructura de los compuestos químicos se puede representar mediante un gráfico, donde los nodos son átomos y los enlaces entre los nodos son enlaces químicos. La Figura 2.3 (b) muestra un diagrama de bola y palo del compuesto químico benceno, que contiene átomos de carbono (negro) e hidrógeno (gris). Una representación gráfica permite determinar qué subestructuras ocurren con frecuencia en un conjunto de compuestos y determinar si la presencia de alguna de estas subestructuras está asociada con la presencia o ausencia de ciertas propiedades químicas, como el punto de fusión o el calor de la formación.



(a) Linked Web pages. (b) Benzene molecule.

Figure 2.3. Different variations of graph data.

Datos ordenados

Para algunos tipos de datos, los atributos tienen relaciones que involucran orden en el tiempo o el espacio.

Datos secuenciales: Los datos secuenciales, también conocidos como datos temporales, se pueden considerar como una extensión de los datos de registro, donde cada registro lleva un tiempo asociado. Considere un conjunto de datos de transacciones minoristas que también almacena el momento en que se realizó la transacción. Esta información de tiempo permite encontrar patrones como "pico de ventas de dulces antes de Halloween". También se puede asociar un tiempo a cada atributo. Por ejemplo, cada registro podría ser el historial de compras de un cliente, con una lista de artículos comprados en diferentes momentos. Usando esta información, es posible encontrar patrones como "las personas que compran reproductores de DVD tienden a comprar DVD en el período inmediatamente posterior a la compra".

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

(a) Sequential transaction data.

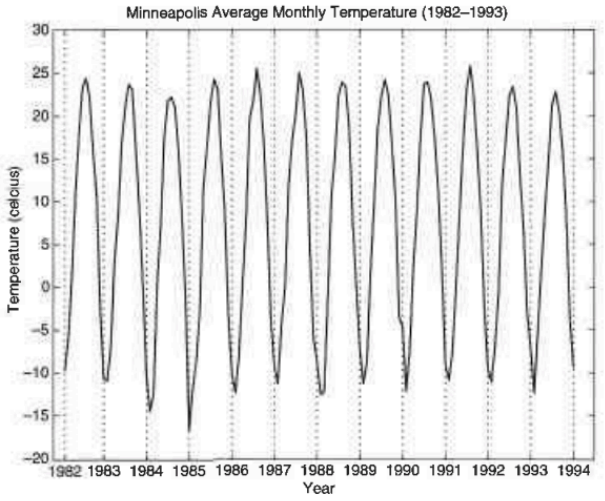
Datos de secuencia: los datos de secuencia consisten en un conjunto de datos que es una secuencia de entidades individuales, como una secuencia de palabras o letras. Es bastante similar a los datos secuenciales, excepto que no hay sellos de tiempo; en cambio, hay posiciones en una secuencia ordenada. Por ejemplo, la información genética de plantas y animales se puede representar en forma de secuencias de nucleótidos que se conocen como genes. Muchos de los problemas asociados con los datos de secuencia genética implican predecir similitudes en la estructura y función de los genes a partir de similitudes en secuencias de nucleótidos. La Figura de la derecha muestra una sección del código genético humano expresado utilizando los cuatro nucleótidos a partir de los cuales se construye todo el ADN: A, T, G y C.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.

Datos de series de tiempo: los datos de series de tiempo son un tipo especial de datos secuenciales en los que cada registro es una serie de tiempo, es decir, una serie de mediciones tomadas a lo largo del tiempo. Por ejemplo, un conjunto de datos financieros puede contener objetos que son series de tiempo de los precios diarios de varias acciones. Como otro ejemplo ver la gráfica de más abajo que muestra una serie de tiempo de la temperatura mensual promedio de Minneapolis durante los años 1982 a 1994. Al trabajar con datos temporales, es importante considerar la autocorrelación temporal; es decir, si dos mediciones están cerca en el tiempo, entonces los valores de esas mediciones son a menudo muy similares.

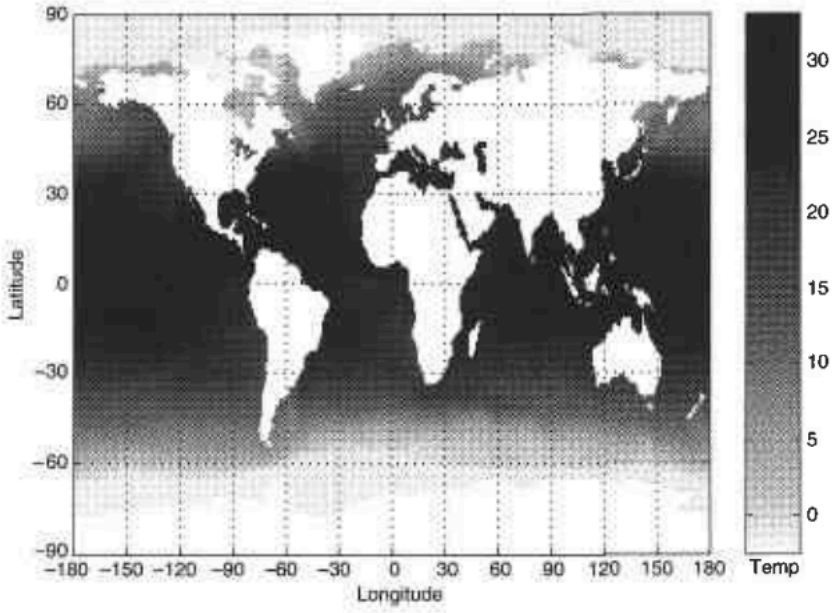
Un ejemplo muy similar es una gráfica que muestre los precios del bitcoin o el ethereum en una serie de tiempo.



(c) Temperature time series.

Datos espaciales: algunos objetos tienen atributos espaciales, como posiciones o áreas, así como otros tipos de atributos. Un ejemplo de datos espaciales son los datos meteorológicos (precipitación, temperatura, presión) que se recopilan para una variedad de ubicaciones geográficas. Un aspecto importante de los datos espaciales es la autocorrelación espacial; es decir, los objetos que están físicamente cerca tienden a ser similares en otras formas también. Por lo tanto, dos puntos en la Tierra que están cerca uno del otro generalmente tienen valores similares para la temperatura y la lluvia.

Ejemplos importantes de datos espaciales son los conjuntos de datos de ciencia e ingeniería que son el resultado de mediciones o resultados de modelos tomados en puntos distribuidos regular o irregularmente en una cuadrícula o malla de dos o tres dimensiones. Por ejemplo, los conjuntos de datos de la ciencia de la Tierra registran la temperatura o la presión medida en puntos (celdas de la cuadrícula) en cuadrículas esféricas de latitud-longitud de varias resoluciones, Como otro ejemplo, en la simulación del flujo de un gas, la velocidad y la dirección del flujo se pueden registrar para cada punto de la cuadrícula en la simulación.



(d) Spatial temperature data.

Calidad de los datos

Las aplicaciones de minería de datos a menudo se aplican a los datos que se recopilaron para otro propósito o para aplicaciones futuras, pero no especificadas. Por esa razón, la minería de datos por lo general no puede aprovechar los beneficios significativos de "abordar los problemas de calidad en la fuente". En contraste, gran parte de las estadísticas se ocupan del diseño de experimentos o encuestas que logran un nivel de calidad de datos pre-especificado. Debido a que la prevención de problemas de calidad de datos no suele ser una opción, la extracción de datos se centra en:

- La detección y corrección de problemas de calidad de datos (a menudo llamado limpieza de datos)
- El uso de algoritmos que pueden tolerar una calidad de datos deficiente.

No es realista esperar que los datos sean perfectos. Puede haber problemas debido a errores humanos, limitaciones de dispositivos de medición o fallas en el proceso de recolección de datos. Pueden faltar valores o incluso objetos de datos completos. En otros casos, puede haber objetos falsos o duplicados; es decir, múltiples objetos de datos que todos corresponden a un único objeto "real". Por ejemplo, puede haber dos registros diferentes para una persona que ha vivido recientemente en dos direcciones diferentes. Incluso si todos los datos están presentes y "parecen estar bien", puede haber inconsistencias: una persona tiene una altura de 2 metros, pero pesa solo 2 kilogramos.

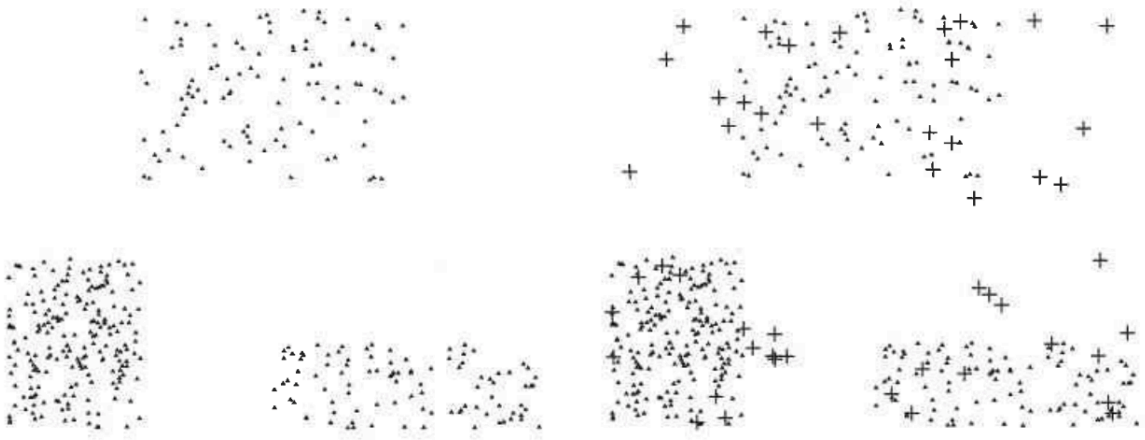
Errores de medición y recopilación de datos

El término error de medición se refiere a cualquier problema que resulte del proceso de medición. Un problema común es que el valor registrado difiere del valor verdadero en cierta medida. Para los atributos continuos, la diferencia numérica del valor medido y verdadero se denomina error. El término error de recopilación de datos se refiere a errores, como omitir objetos de datos o valores de atributos, o incluir de manera inapropiada un objeto de datos. Por ejemplo, un estudio de animales de una cierta especie podría incluir animales de una especie relacionada que son similares en apariencia a las especies de interés. Tanto los errores de medición como los errores de recolección de datos pueden ser sistemáticos o aleatorios.

Solo consideraremos tipos generales de errores. Dentro de dominios particulares, hay ciertos tipos de errores de datos que son comunes y, por lo tanto, existen técnicas bien desarrolladas para detectar y / o corregir estos errores. Por ejemplo, los errores de teclado son comunes cuando los datos se ingresan manualmente y, como resultado, muchos programas de ingreso de datos tienen técnicas para detectar y, con la intervención humana, corregir dichos errores.

Ruido y artefactos

El ruido es el componente aleatorio de un error de medición. Puede implicar la distorsión de un valor o la adición de objetos falsos. La Figura 2.5 muestra una serie de tiempo antes y después de que haya sido interrumpida por ruido aleatorio. Si se añadiera un poco más de ruido a la serie temporal, se perdería su forma. La siguiente figura muestra un conjunto de puntos de datos antes y después de que se hayan agregado algunos puntos de ruido (indicados por '+'). Observe que algunos de los puntos de ruido se entremezclan con los puntos sin ruido.



(a) Three groups of points. (b) With noise points (+) added.

Figure 2.6. Noise in a spatial context.

Precisión, sesgo y exactitud

En estadística y ciencia experimental, la calidad del proceso de medición y los datos resultantes se miden por precisión y sesgo. Proporcionamos las definiciones estándar, seguido de una breve discusión. Para las siguientes definiciones, asumimos que realizamos mediciones repetidas de la misma cantidad subyacente y usamos este conjunto de valores para calcular un valor medio (promedio) que sirve como nuestra estimación del valor verdadero.

Precisión: La cercanía de mediciones repetidas (de la misma cantidad) entre sí.

Sesgo: Una variación sistemática de las medidas de la cantidad que se mide.

Exactitud: La proximidad de las mediciones al valor real de la cantidad que se mide.

Algunas veces se pasan por alto cuestiones como los dígitos significativos, la precisión, el sesgo y la exactitud, pero son importantes para la minería de datos, así como para las estadísticas y la ciencia. Muchas veces, los conjuntos de datos no vienen con información sobre la precisión de los datos y, además, los programas utilizados para el análisis devuelven resultados sin tal información. No obstante, sin comprender la exactitud de los datos y los resultados, un analista corre el riesgo de cometer errores graves de análisis de datos.

Valores atípicos:

Los valores atípicos son objetos de datos que, en cierto sentido, tienen características que son diferentes de la mayoría de los otros objetos de datos en el conjunto de datos, o valores de un atributo que son inusuales con respecto a los valores típicos para ese atributo. Alternativamente, podemos hablar de objetos o valores anómalos. Hay un margen de maniobra considerable en la definición de un valor atípico, y las comunidades de minería de datos y estadísticas han propuesto muchas definiciones diferentes. Además, es importante distinguir entre las nociones de ruido y valores atípicos. Los valores atípicos pueden ser

objetos o valores de datos legítimos. Por lo tanto, a diferencia del ruido, los valores atípicos a veces pueden ser interesantes. En la detección de fraudes e intrusiones en la red, por ejemplo, el objetivo es encontrar objetos o eventos inusuales entre una gran cantidad de objetos normales. El capítulo 10 analiza la detección de anomalías con más detalle.

Valores faltantes

No es raro que a un objeto le falten uno o más valores de atributo. En algunos casos, la información no fue recopilada; por ejemplo, algunas personas se niegan a dar su edad o peso. En otros casos, algunos atributos no son aplicables a todos los objetos; Por ejemplo, a menudo, los formularios tienen partes condicionales que se completan solo cuando una persona responde a una pregunta anterior de cierta manera, pero por simplicidad, todos los campos se almacenan. Independientemente, los valores faltantes deben tenerse en cuenta durante el análisis de los datos.

Existen varias estrategias (y variaciones en estas estrategias) para tratar los datos faltantes, cada uno de los cuales puede ser apropiado en ciertas circunstancias. Estas estrategias se enumeran a continuación, junto con una indicación de sus ventajas y desventajas.

Eliminar objetos de datos o atributos

Una estrategia simple y efectiva es eliminar objetos con valores faltantes. Sin embargo, incluso un objeto de datos parcialmente especificado contiene cierta información, y si a muchos objetos les faltan valores, un análisis confiable puede ser difícil o imposible. No obstante, si un conjunto de datos tiene solo unos pocos objetos que tienen valores faltantes, puede ser conveniente omitirlos. Una estrategia relacionada es eliminar los atributos que tienen valores faltantes. Sin embargo, esto debe hacerse con precaución, ya que los atributos eliminados pueden ser los que son críticos para el análisis.

Estimar valores faltantes

A veces, los datos faltantes se pueden estimar de manera confiable. Por ejemplo, considere una serie de tiempo que cambie de una manera razonablemente suave, pero que tenga unos pocos valores perdidos, muy dispersos. En tales casos, los valores faltantes pueden estimarse (interpolados) bV utilizando los valores restantes. Como otro ejemplo, considere un conjunto de datos que tiene muchos puntos de datos similares. En esta situación, los valores de atributo de los puntos más cercanos al punto con el valor faltante se usan a menudo para estimar el valor faltante. Si el atributo es continuo, se usa el valor de atributo promedio de los vecinos más cercanos; Si el atributo es categórico, entonces se puede tomar el valor de atributo más común. Para una ilustración concreta, considere las medidas de precipitación registradas por las estaciones terrestres. Para áreas que no contienen una estación terrestre, la precipitación puede estimarse utilizando valores observados en estaciones terrestres cercanas.

Ignorar el valor faltante durante el análisis

Muchos enfoques de minería de datos pueden modificarse para ignorar los valores faltantes. Por ejemplo, suponga que los objetos se agrupan y que se debe calcular la similitud entre pares de objetos de datos. Si uno o ambos objetos de un par tienen valores faltantes para

algunos atributos, entonces la similitud se puede calcular utilizando solo los atributos que no tienen valores faltantes. Es cierto que la similitud solo será aproximada, pero a menos que el número total de atributos sea pequeño o el número de valores faltantes sea alto, este grado de inexactitud puede no importar mucho. Del mismo modo, muchos esquemas de clasificación pueden modificarse para trabajar con valores perdidos.

Valores Inconsistentes

Los datos pueden contener valores inconsistentes. Considere un campo de dirección, donde se enumeran tanto un código postal como una ciudad, pero el área del código postal especificado no está contenida en esa ciudad. Puede ser que la persona que ingresa esta información haya traspuesto dos dígitos, o tal vez un dígito se haya leído mal cuando la información se ha escaneado de una forma manuscrita. Independientemente de la causa de los valores inconsistentes, es importante detectar y, si es posible, corregir tales problemas.

Algunos tipos de inconsistencias son fáciles de detectar. Por ejemplo, la altura de una persona no debe ser negativa. En otros casos, puede ser necesario consultar una fuente externa de información. Por ejemplo, cuando una compañía de seguros procesa los reclamos de reembolso, verifica los nombres y direcciones en los formularios de reembolso contra una base de datos de sus clientes.

Una vez que se ha detectado una inconsistencia, a veces es posible corregir los datos. Un código de producto puede tener dígitos de "verificación", o puede ser posible volver a verificar un código de producto contra una lista de códigos de producto conocidos, y luego corregir el código si es incorrecto, pero cerca de un código conocido. La corrección de una inconsistencia requiere información adicional o redundante.

Datos duplicados

Un conjunto de datos puede incluir objetos de datos que son duplicados, o casi duplicados, entre sí. Muchas personas reciben correos duplicados porque aparecen en una base de datos varias veces con nombres ligeramente diferentes. Para detectar y eliminar tales duplicados, se deben abordar dos problemas principales. Primero, si hay dos objetos que realmente representan un solo objeto, entonces los valores de los atributos correspondientes pueden diferir, y estos valores inconsistentes deben resolverse. En segundo lugar, se debe tener cuidado para evitar la combinación accidental de objetos de datos similares, pero no duplicados, como dos personas distintas con nombres idénticos. El término deduplicación se usa a menudo para referirse al proceso de tratar estos problemas.

En algunos casos, dos o más objetos son idénticos con respecto a los atributos medidos por la base de datos, pero aún representan diferentes objetos. Aquí, los duplicados son legítimos, pero aún pueden causar problemas para algunos algoritmos si la posibilidad de objetos idénticos no se tiene en cuenta específicamente en su diseño.

Problemas relacionados con las aplicaciones

Los problemas de calidad de los datos también se pueden considerar desde el punto de vista de la aplicación, como se expresa en la declaración "los datos son de alta calidad si son adecuados para el uso previsto". Este enfoque de la calidad de los datos ha demostrado ser bastante útil, particularmente en los negocios y la industria. Un punto de vista similar también está presente en las estadísticas y las ciencias experimentales, con su énfasis en el diseño cuidadoso de experimentos para recopilar los datos relevantes para una hipótesis específica.

Al igual que con los problemas de calidad en el nivel de medición y recopilación de datos, hay muchos problemas que son específicos de aplicaciones y campos específicos. Una vez más, consideramos sólo algunos de los problemas generales.

Conocimiento sobre los datos.

Idealmente, los conjuntos de datos van acompañados de documentación que describe diferentes aspectos de los datos; La calidad de esta documentación puede ayudar u obstaculizar el análisis posterior. Por ejemplo, si la documentación identifica varios atributos como fuertemente relacionados, es probable que estos atributos proporcionen información altamente redundante, y podemos decidir mantener solo uno. (Considere el impuesto a las ventas y el precio de compra). Sin embargo, si la documentación es deficiente y no nos indica, por ejemplo, que los valores que faltan para un campo en particular se indican con un -9999, entonces nuestro análisis de los datos puede ser erróneo. . Otras características importantes son la precisión de los datos, el tipo de características (nominal, ordinal, intervalo, relación), la escala de medición (por ejemplo, metros o pies para la longitud) y el origen de los datos.