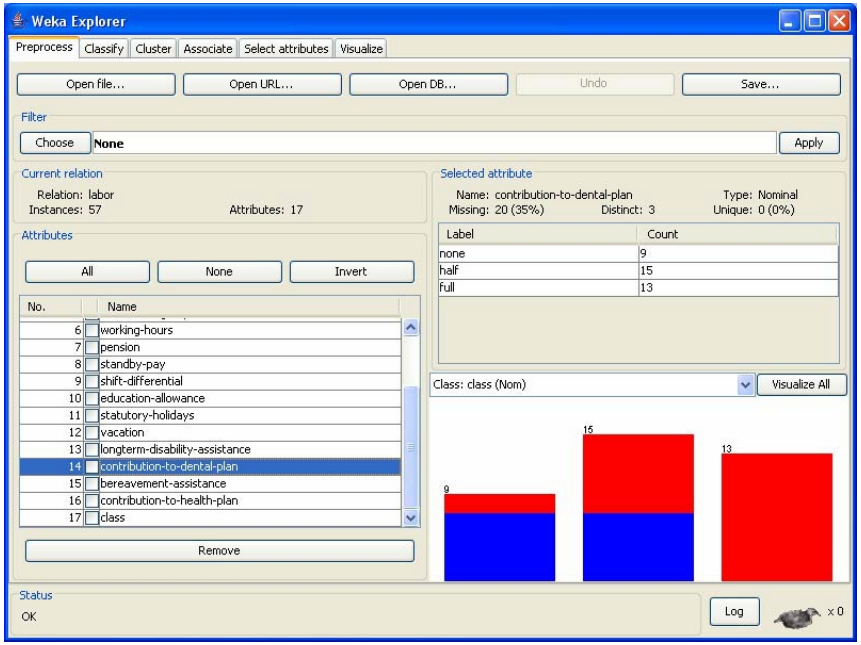
Repositorios de Datos Avanzados

- Bases de datos Orientadas a Objetos
- Bases de datos Objeto-Relacionales
- Bases de datos Espaciales
- Bases de datos Temporales y de Series de Tiempo
- Bases de datos de Texto
- Bases de datos Multimedia
- Bases de datos Heterogéneas
- WorldWideWeb

Selección, Limpieza y Transformación de Datos

Limpieza y selección de datos

El primer paso en la limpieza de datos consiste en la elaboración de un resumen de características.



Limpieza y selección de datos

Datos Nominales.

Debemos analizar con detalle cada uno de los atributos Nominales. Podemos detectar:

- Valores redundantes

(Hombre,Varón)

- Valores despreciables

(agrupar valores como otros)

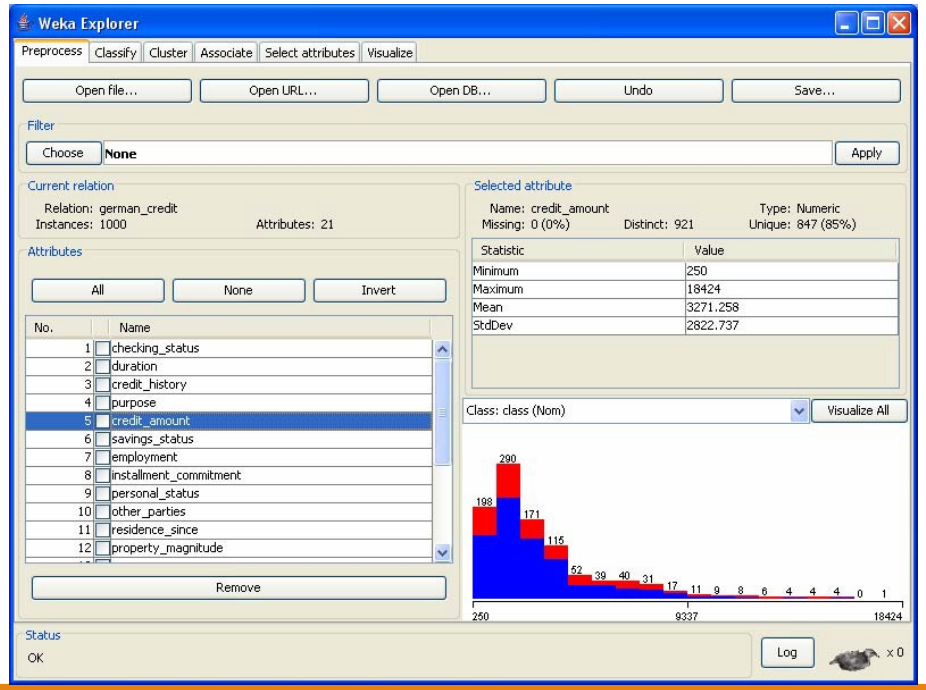
Limpieza y selección de datos

Atributos Numéricos.

Debemos analizar con detalle cada uno de los atributos. Podemos detectar:

- Valores anómalos

- Distribuciones en los datos



Limpieza (data cleansing) y selección de datos

Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (selección o criba).

Métodos estadísticos casi exclusivamente.

- histogramas (detección de datos anómalos).
- selección de datos (ya sea verticalmente, eliminando atributos u horizontalmente, eliminando tuplas).
- redefinición de atributos (agrupación o separación).

Limpieza (data cleansing) y selección de datos

Muy relacionado con la disciplina de “Calidad de Datos”.

Mejor si se tienen metadatos acerca de la calidad de datos (frec. de uso, etc.)

Datos Anómalos - Acciones

- ignorar: algunos algoritmos son robustos a datos anómalos (p.ej. Árboles)
- filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna discreta diciendo si el valor era normal o outlier (por encima o por debajo).
- filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.

Datos Anómalos - Acciones

- reemplazar el valor: por el valor ‘nulo’ si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.

- discretizar: transformar un valor continuo en uno discreto (p.ej. muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en ‘muy alto’ o ‘muy bajo’ sin mayores problemas.

Datos Faltantes - Acciones

- ignorar
- filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna, es reemplazarla por una columna booleana diciendo si el valor existía o no.
- filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.

Datos Faltantes - Acciones

- reemplazar el valor: por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.
- segmentar: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Inconsistencias

Un problema grave que afecta a varios métodos de aprendizaje predictivo son los registros inconsistentes, es decir, dos o más registros con los mismo valores en los atributos, pero diferente valor en el atributo clase.

Algunas técnicas no soportan las inconsistencias en los datos. Por lo que se deben eliminar unificando (siempre que se pueda) los registros en una única clase.

Transformación de Atributos

La transformación de datos engloba, en realidad, cualquier proceso que modifique la forma de los datos.

Por transformación entendemos aquellas técnicas que transforman un conjunto de atributos en otros, o bien derivan nuevos atributos, o bien cambian el tipo o el rango.

La selección de atributos (eliminar los menos relevantes) en realidad no transforma atributos y, en consecuencia, no entra en este grupo de técnicas.

Reducción de Dimensionalidad

Si tenemos muchas dimensiones (atributos) respecto a la cantidad de instancias, pueden existir demasiados grados de libertad, por lo que los patrones extraídos pueden ser poco robustos.

Este problema se conoce popularmente como “la maldición de la dimensionalidad” (“the curse of dimensionality”). Una manera de intentar resolver este problema es mediante la reducción de dimensiones.

La reducción se puede realizar por selección de un subconjunto de atributos, o bien la sustitución del conjunto de atributos iniciales por otros diferentes.

Análisis de Componentes Principales

La técnica más conocida para reducir la dimensionalidad por transformación se denomina “análisis de componentes principales” (“principal component analysis”), PCA.

PCA transforma los m atributos originales en otro conjunto de atributos p donde $p \leq m$.

Este proceso se puede ver geométricamente como un cambio de ejes en la representación (proyección).

Los nuevos atributos se generan de tal manera que son independientes entre sí y, además, los primeros tienen más relevancia (más contenido informacional) que los últimos.

Aumento de Dimensionalidad

En ocasiones añadir atributos nuevos puede mejorar el proceso de aprendizaje.

Técnicas de generación de nuevos atributos:

- Numéricos: generalmente, operaciones matemáticas básicas de uno o más argumentos
- Nominales: Operaciones lógicas: conjunción, disyunción, negación, implicación, condiciones M-de-N (M-de-N es cierto si y sólo si al menos M de las N condiciones son ciertas), igualdad o desigualdad

El conocimiento del dominio es el factor que más determina la creación de buenos atributos derivados.

Discretización de Atributos

La discretización, o cuantización ("binning") es la conversión de un valor numérico en un valor nominal ordenado.

La discretización se debe realizar cuando:

- El error en la medida puede ser grande
- Existen umbrales significativos (p.e. notas)
- En ciertas zonas el rango de valores es más importante que en otras (interpretación no lineal)
- Aplicar ciertas tareas de MD que sólo soportan atributos nominales (p.e. reglas de asociación)

Discretización de Atributos

La discretización más sencilla (simple binning) es aquella que realiza intervalos del mismo tamaño y utilizando el mínimo y el máximo como referencia.

Otra técnica de discretización sencilla es intentar obtener intervalos con el mismo número de registros (Equal-frequency binning) .

Sin embargo estas técnicas de discretización ignoran la clase, por lo que pueden dar lugar a intervalos no adecuados.

Numerización de Atributos

La numerización es el proceso inverso a la discretización, es decir, convertir un atributo nominal en numérico.

La discretización se debe realizar cuando se quieren aplicar ciertas técnicas de MD que sólo soportan atributos numéricos (p.e. Regresión, métodos basados en distancias).

Numerización de Atributos

numerización "1 a n": Si una variable nominal x tiene posibles valores creamos n variables numéricas, con valores 0 o 1 dependiendo de si la variable nominal toma ese valor o no.

Podemos también prescindir del último atributo pues es dependiente del resto (numerización "1 a n-1").

numerización "1 a 1": Se aplica si existe un cierto orden o magnitud en los valores del atributo nominal. Por ejemplo, si tenemos categorías del estilo {niño, joven, adulto, anciano} podemos numerar los valores de 1 a 4.

Normalización de Atributos

Algunos métodos de aprendizaje funcionan mejor con los atributos normalizados. Por ejemplo, los métodos basados en distancias.

La normalización más común es la normalización lineal uniforme:

$$v' = (v-min) / (max-min)$$

Esta normalización es muy sensible a la presencia de valores anómalos (outliers).

Métodos de 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 y, en cierto modo, independiente de él.

- 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).

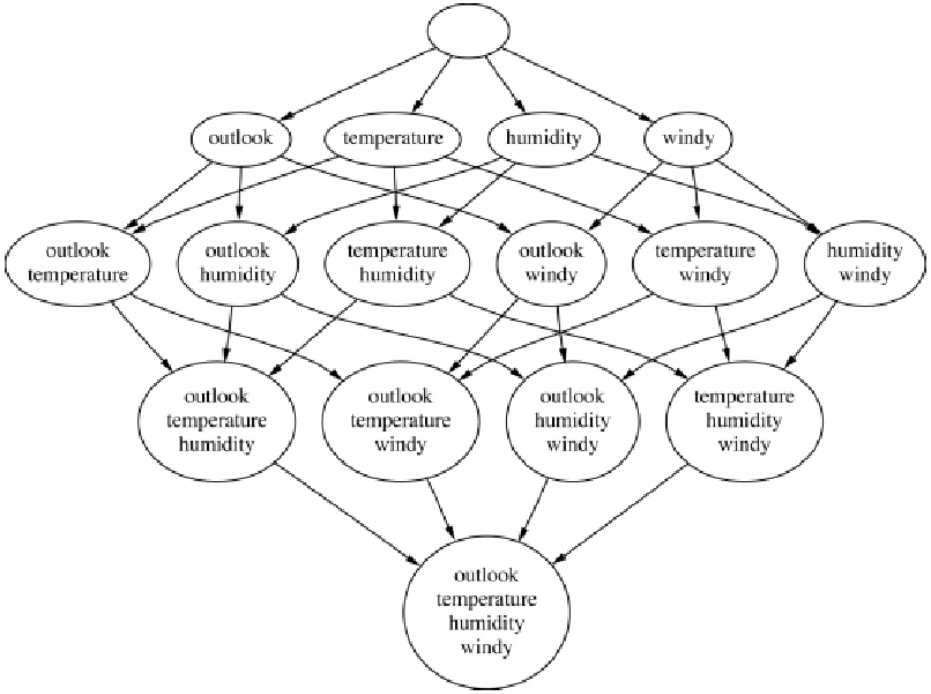
Métodos de Selección de Características

En cualquier caso se emplea una estrategia iterativa, es decir, se van eliminando atributos y se va observando el resultado. Se van recuperando o eliminando más atributos de una manera iterativa, hasta que se obtiene una combinación que maximiza la calidad.

De acuerdo con la medida de evaluación, estrategia y dirección de búsqueda, podemos establecer taxonomías más refinadas.

Métodos de Selección de Características

El número de subconjuntos de atributos crece exponencialmente con respecto al número de atributos.



Métodos de Selección de Características

Medida de Evaluación

- Clásicas: ganancia de información, o medidas de dependencia entre características.
- Acierto: Precisión u cualquier otra medida de evaluación de calidad medido sobre un conjunto de test.
- Consistencia: medidas que miden el grado de inconsistencias (registros iguales salvo en la clase) en el conjunto de datos.

Métodos de Selección de Características

Estrategia de Búsqueda

- Completa: Se cubren todas las combinaciones posibles de selección.
- Heurística: reduce el número de combinaciones a evaluar basándose en algún tipo de información.
- No determinista (estocástico): basada en algoritmos de búsqueda globales. Intentan evitar el problema de mínimos locales.

Métodos de Selección de Características

Dirección de Búsqueda

- Forward: Empezando con el mejor atributo y añadir el atributo que dé mayor calidad de selección con dos atributos, y así hasta que no se mejore la calidad o se llegue al número deseado de atributos.
- Backward: Se inicia el proceso con todos los atributos eliminando uno a uno el menos relevante.
- Aleatoria: Se producen patrones de búsqueda mediante la creación de conjuntos de manera aleatoria.

Métodos de Selección de Características

Análisis Correlacional: Una técnica sencilla que consiste en utilizar una matriz de correlaciones.

	Pulsaciones	Alcoholismo	Tabaquismo	Colesterol	Obesidad	Tensión	Edad
Pulsaciones		0,27	0,32	0,4	0,39	0,23	-0,15
Alcoholismo	0,15		0,32	0,27	0,58	0,23	-0,22
Tabaquismo	-0,02	0,72		0,52	0,58	0,39	-0,12
Colesterol	0,42	0,56	0,67		0,27	0,4	0,45
Obesidad	0,34	0,22	0,67	0,72		0,32	0,21
Tensión	0,63	0,22	0,56	0,72	0,43		-0,08
Edad	0,63	0,34	0,42	-0,02	0,15	0,12	

Métodos de Selección de Características

Análisis por modelo lineal: Queremos obtener un modelo de predicción de azúcar en la sangre a partir de los datos anteriores.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

ATRIBUTO	EDAD	TENSIÓN	OBES.	COLEST.	TABAQ.	ALCOHOL.	PULS.	HIERRO
Azúcar	2,23	-1,63	3,23	0,42	-0,12	2,23	0	-3,01

Esto, en realidad, es sólo el principio de múltiples técnicas del análisis multivariante. Si quisiéramos saber si el colesterol, el tabaquismo y las pulsaciones no influyen en el azúcar y, son, por tanto, descartables, deberíamos usar, por ejemplo, el Análisis de la Varianza (ANOVA).