SELECCIÓN DE VARIABLES

GARCÍA VEIGA, MARIAM (USC)
LADO GONZÁLEZ, IGNACIO (USC)
LEYENDA RODRÍGUEZ, MARÍA (USC)
LÓPEZ VEIGA, DAVID (USC)

27 de Noviembre de 2009

ÍNDICE

1. INTRODUCCIÓN 4

2. ALGORITMOS DE PONDERACIÓN BINARIA 6

2.1. Cfs Subset Eval 6

2.1.1. Método de búsqueda: 6

2.2. Classifier Subset Eval 6

2.2.1. Método de búsqueda: 6

2.3. Consistency Subset Eval 6

2.3.1. Método de búsqueda: 6

2.4. Cost Sensitive Subset Eval 6

2.4.1. Método de búsqueda: Greedy Stepwise 6

2.5. Fitered Subset Eval 6

2.5.1. Método de búsqueda: 6

2.6. Wrapper Subset Eval 6

2.6.1. Método de búsqueda: 6

2.7. Symmetrical Uncert Attribute Set Eval 6

2.7.1. Método de búsqueda: 6

3. ALGORITMOS DE PONDERACIÓN CONTÍNUA 7

3.1. Método de búsqueda: RANKER 7

3.2. Chi Square Attribute Eval 7

3.3. Cost Sensitive Attribute Eval 7

3.4. Filtered Attribute Eval 7

3.5. Gain Ratio Attribute Eval 7

3.6. Info Gain Attribute Eval 7

3.7. OneR Atttribute Eval 7

3.8. RelievefF Attribute Eval 7

3.9. Symmetrical Uncert Attribute Eval 9

4. LATENT SEMANTIC ANALYSIS 10

5. COMPONENTES PRINCIPALES 11

6. PROBLEMA 1: ENCUESTA SOBRE ACCIDENTES EN EMPRESAS MINERAS 13

6.1. APLICACIÓN: ALGORITMOS DE PONDERACIÓN BINARIA 13

6.2. APLICACIÓN: ALGORITMOS DE PONDERACIÓN CONTÍNUA 13

6.3. APLICACIÓN: LATENT SEMANTIC ANÁLISIS 14

6.4. APLICACIÓN COMPONENTES PRINCIPALES 14
7. PROBLEMA 2: CREDIT-SCORING 15
7.1. APLICACIÓN: ALGORITMOS DE PONDERACIÓN BINARIA 16
7.2. APLICACIÓN: ALGORITMOS DE PONDERACIÓN CONTÍNUA 16
7.3. APLICACIÓN: LATENT SEMANTIC ANÁLISIS 16
7.4. APLICACIÓN COMPONENTES PRINCIPALES 16

1. INTRODUCCIÓN

La selección de variables es un problema muy estudiado, aunque fundamentalmente abierto. Las fuertes interacciones entre las variables y la presencia de variables irrelevantes, redundantes, el ruido en la muestra, etc., dificultan aún más el problema.

Vamos a estudiar el problema de selección de variables a través de algoritmos. El problema de desarrollar un ASV es básicamente uno de búsqueda en un espacio de estados. Cada *estado* representa un subconjunto de variables ponderadas; el objetivo es encontrar el estado con la mejor medida de evaluación. El número de subconjuntos potenciales a evaluar es 2n en caso que la ponderación sea binaria.

Existen 2 tipos de ASV

- Algoritmos que proporcionan un orden lineal de las variables (ponderación continua).
- Algoritmos que obtienen un subconjunto del conjunto original (ponderación binaria).
-

Vamos a estudiar la selección de atributos utilizando la herramienta Weka. Para ello vamos a usar dos conjuntos de datos:"Encuesta de accidentes.xls" y "credit.xls".

El sistema Weka incorpora una gran cantidad de métodos para estudiar la **relevancia de atributos** y realizar una **selección automática de los mismos**. Estos métodos, están dentro de la entorno *Explorer* en la sección *Select Attributes*. Esta sección permite automatizar la búsqueda de subconjuntos de atributos más apropiados para "explicar" un atributo objetivo, en un sentido de clasificación supervisada: permite explorar qué subconjuntos de atributos son los que mejor pueden clasificar la clase de la instancia.

La selección supervisada de atributos tiene dos componentes:

- **Método de búsqueda**(Search Method): es la forma de realizar la búsqueda de conjuntos. Como la evaluación exhaustiva de todos los subconjuntos es un problema combinatorio inabordable en cuanto crece el número de atributos, aparecen estrategias que permiten realizar la búsqueda de forma eficiente
    - "**SubSetEval"**: Evaluadores de conjuntos o selectores. Estos necesitan elegir un método o estrategia de búsqueda de los subconjuntos .

- ***Método de Evaluación*** (**Attribute Evaluator):** es la función que determina la calidad del conjunto de atributos para discriminar la clase.
    - **"AttributeEval"**: Porteadores de atributos. Estos solo pueden combinarse con un "Ranker" ya que no seleccionan atributos sino que solo los ordenan por relevancia.

Dentro los método de evaluación podemos distinguir dos tipos: Los métodos que directamente utilizan un clasificador específico para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador (métodos “wraper”) y los que no.

- **Métodos "wrapper",** porque "envuelven" al clasificador para explorar la mejor selección de atributos que optimiza sus prestaciones, son muy costosos porque necesitan un proceso completo de entrenamiento y evaluación en cada paso de búsqueda.
- Métodos como el método **"CfsSubsetEval",** que calcula la correlación de la clase con cada atributo, y eliminan atributos que tienen una correlación muy alta como atributos redundantes.

Hay diferentes métodos de búsqueda de las variables más influyentes, como son:

- "**ForwardSelection",** que es un método de búsqueda muy rápido que subóptima en escalada, donde elije primero el mejor atributo, después añade el siguiente atributo que más aporta y continua así hasta llegar a la situación en la que añadir un nuevo atributo empeora la situación.
- "**BestSearch",** que permite buscar interacciones entre atributos más complejas que el análisis incremental anterior. Este método va analizando lo que mejora y empeora un grupo de atributos al añadir elementos, con la posibilidad de hacer retrocesos para explorar con más detalle.
- **"ExhaustiveSearch"** simplemente enumera todas las posibilidades y las evalúa para seleccionar la mejor.

Por otro lado, en la configuración del problema debemos seleccionar que atributo objetivo se utiliza para la selección supervisada, en la ventana de selección y determinar si la evaluación se realizará con todas las instancias disponibles o mediante validación cruzada.

2. ALGORITMOS DE PONDERACIÓN BINARIA

2.1. Cfs Subset Eval

2.1.1. Método de búsqueda:

2.2. Classifier Subset Eval

2.2.1. Método de búsqueda:

2.3. Consistency Subset Eval

2.3.1. Método de búsqueda:

2.4. Cost Sensitive Subset Eval

2.4.1. Método de búsqueda: Greedy Stepwise

2.5. Fitered Subset Eval

2.5.1. Método de búsqueda:

2.6. Wrapper Subset Eval

2.6.1. Método de búsqueda:

2.7. Symmetrical Uncert Attribute Set Eval

2.7.1. Método de búsqueda:

3. ALGORITMOS DE PONDERACIÓN CONTÍNUA

3.1. Método de búsqueda: RANKER

Es una función de Weka que determina el rango de los atributos(variables) por sus evaluaciones individididuales. Se usa en conjunción con los evaluadores que ordenan los atributos por relevancia. (ReliefF, GainRatio, Entropy etc). Es decir, se usa junto con los algoritmos de ponderación contínua.

3.2. Chi Square Attribute Eval

3.3. Cost Sensitive Attribute Eval

3.4. Filtered Attribute Eval

3.5. Gain Ratio Attribute Eval

3.6. Info Gain Attribute Eval

3.7. OneR Atttribute Eval

Es un algoritmo implementado en Weka que evalúa el valor de las variables usando un clasificador OneR.

OneR, es uno de los clasificadores más sencillos y rápidos, aunque en ocasiones, sus resultados son sorprendentemente buenos en comparación con algoritmos mucho más complejos. Este clasificador, simplemente selecciona el atributo que mejor "explica" la clase de salida. Si hay atributos numéricos, busca los umbrales para hacer reglas con mejor tasa de aciertos.

3.8. RelievefF Attribute Eval

Evalúa el valor de un atributo por muestreo repetidamente y considerando el valor de la atributo de la muestra de entrenamiento más cercana de la misma y clase diferente. Puede funcionar en ambas clases de datos discreta y continua. Utiliza el algoritmo RELIEF el cual es un algoritmo estadístico de selección de variables que usa muestras de entrenamiento para asignar peso relevante a cada característica.

Relief es un algoritmo de selección de variables predictoras inspirado en el conocimiento. Dado un conjunto de datos de entrenamiento S, muestra de tamaño m, y un umbral de relevancia ζ, Relief detecta esas variables predictoras que son estadísticamente relevantes.

Sea

- S denota una colección de datos de entrenamiento de tamaño n.
- F es una colección de variables predictoras dada

$\{f_1, f_2, \ldots, f_p\}$.

- X es denotado por un vector p-dimensional $(x_1, x_2, \ldots, x_p)$

Donde x_j denota el valor de la variable predictora f_j de X.

- ζ codifica un umbral de relevancia ($0 \leq t \leq 1$). Se asume que la escala de las variables predictoras es nominal o numérica (entera o real). Los diferentes valores de las variables predictoras entre dos instantes X e Y son definidos por la siguiente función diff.

- Cuando x_k e y_k son nominales:

$$\text{diff}(x_k, y_k) = \begin{cases} 0 & \text{si } x_k \text{ e } y_k \text{ son la misma} \\ 1 & \text{si } x_k \text{ e } y_k \text{ son diferentes} \end{cases}$$

- Cuando x_k e y_k son numéricas:

  - $\text{diff}(x_k, y_k) = (x_k - y_k)/nu_k$ dónde nu_k es una normalización unidad para normalizar los valores de diff en el intervalo [0,1]

Relief escoge una muestra compuesta por m tripletes de un dato X, sus datos Near-hit y Near-miss dato.

- Near-hit: Si pertenece a los vecinos cercanos de X y tiene la misma categoría que X.
- Near-miss: Si pertenece a los vecinos cercanos de X pero no tiene la misma categoría que X.

Relief usa la distancia Euclidea p-dimensional para seleccionar Near-hit y Near-miss. Relief llama a una rutina para actualizar el peso del vector de las variables predictoras, W, para todos los tripletes y determinar el promedio del peso de la relevancia del vector de variables predictoras.

Relief selecciona las variables en las que el peso medio de relevancia ('nivel de relevancia') está por encima del umbral ζ .

Relief es válido solo cuando:

- El nivel de relevancia es alto para las variables relevantes y bajo para las variables irrelevantes.
- ζ puede ser escogido para retener las variables relevantes y descartar las irrelevantes.

El análisis teórico muestra que:

- La relevancia es positiva cuando la variable es relevante y próxima a cero o negativa cuando es irrelevante.
- Un método estadístico de intervalos estimados, puede ser usado para determinar el valor de ζ

La complejidad de Relief es θ(pmn) porque calcula la distancia entre X y cada uno de los n datos, tomando θ(p) veces, para determinar su Near-miss and Near-hit dentro de un bucle iterativo m veces. m es una cte que afecta a la exactitud de los niveles de relevancia. Luego, m es escogido

independientemente de p y n, la complejidad está en θ(pn). De este modo el algoritmo puede seleccionar estadísticamente las variables relevantes en tiempo lineal en términos del número de variables y el número de datos de entrenamiento.

3.9. Symmetrical Uncert Attribute Eval

4. LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis (LSA) es una técnica natural del procesamiento del lenguaje, en particular en vectores semánticos, que analiza las relaciones entre una colección de documentos y las palabras que ellos contienen por producir una colección de conceptos relacionados a los documentos y palabras.

LSA puede usar una matriz palabra-documento la cual describe cuantas veces aparecen las palabras en los documentos; es una matriz dónde las filas corresponden a las palabras y las columnas a los documentos.

Esta matriz también es común en los modelos semánticos estándar, sin embargo no es necesario expresarla explícitamente como una matriz, dado que las propiedades matemáticas de las matrices no son usadas.

LSA transforma la matriz de sucesos en una relación entre palabras y algunos conceptos, y una relación entre esos conceptos y los documentos. Así de esta manera, las palabras y los documentos están directamente relacionados a través de los conceptos.

Este nuevo espacio de conceptos puede ser usado para:

- Comparar los documentos en el espacio conceptual
- Encontrar similares documentos a través del lenguaje, después de analizar un conjunto base de documentos traducidos
- Encontrar relaciones entre palabras (sinonimia y polisemia)
- Proporciona una búsqueda de los términos, los traduce en el espacio conceptual, y encuentra documentos parecido.

Después de la construcción de la matriz de sucesos, LSA encuentra un menor rango aproximado a la matriz palabra-documento. Puede haber varias razones para esta aproximación:

- La original matriz palabra-documento es presuntamente grande para el cálculo; en este caso, la aproximación de la matriz con menos rango es interpretado como una aproximación.
- La original matriz palabra-documento tiene demasiado ruido(anécdotas, ejemplos…). En este caso, la aproximación es interpretada como una matriz “poco ruidosa”(mejor que la original).
- La matriz palabra-documento original es supuesta demasiado escasa en relación con la matriz de documento término "verdadera". La matriz original pone en una lista sólo las palabras en cada documento, mientras que nosotros podríamos estar interesados en todas las palabras relacionadas con cada documento - generalmente una colección mucho más de grande debido a la sinonimia.

5. COMPONENTES PRINCIPALES

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Las nuevas componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá, pues, que estudiar tanto el signo como la magnitud de las correlaciones).

El análisis de componentes principales consta de las siguientes fases:

- **Análisis de la matriz de correlaciones**
  - Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.
- **Selección de los factores**
  - La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente.
- **Análisis de la matriz factorial**
  - Una vez seleccionados los componentes principales, se representan en forma de matriz. Cada elemento de ésta representa los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales). La matriz tendrá tantas columnas como componentes principales y tantas filas como variables.
- **Interpretación de los factores**
  - Para que un factor sea fácilmente interpretable debe tener las siguientes características, que son difíciles de conseguir:
    - Los coeficientes factoriales deben ser próximos a 1.
    - Una variable debe tener coeficientes elevados sólo con un factor.
    - No deben existir factores con coeficientes similares.
- **Cálculo de las puntuaciones factoriales**
  - Son las puntuaciones que tienen los componentes principales para cada caso, que nos permitirán su representación gráfica

$$X_{ij} = \Sigma a_{is} Z_{sk\ ;\ S=} 1,\ldots\ldots,k$$

Dónde a son los coeficientes y Z son los valores estandarizados que tienen las variables en cada uno de los sujetos de la muestra.

El objetivo de PCA es encontrar una nueva colección de atributos(esta nueva colección de atributos se denomina por componentes principales, PCs) que verifican las siguientes propiedades: Las PCs son

- Combinaciones lineales de los atributos originales
- Ortogonales entre si
- Capturan la máxima cantidad de variabilidad de los datos.

A menudo, la variabilidad de los datos puede ser capturada por un número relativamente pequeño de PCs, por consiguiente, PCA puede dar como resultado datos con poca dimensión con menos ruido que el modelo original.

PCA depende de la escala de los datos, y por lo tanto los resultados a veces no son concluyentes. Además , las componentes principales no son siempre fáciles para hacer de interpretar.

Además de obtener una nueva colección de variables, PCA también es útil en un problema para obtener mejoras en la clasificación.

Veamos las diferencias de tres variantes de la computación de PCA con el objetivo de obtener mejoras en la clasificación:

> Para todos los subconjuntos basados en PCA primero realizamos un cambio en la media de todos los rasgos tal que la media se hace 0. Denotamos la matriz resultante como M.

- PCA1: Los autovalores y los vectores propios son calculados usando la covarianza de la matriz M. Los nuevos valores del atributo son luego calculados al multiplicar M con los vectores propios de Cov(M).
- PCA2: Los autovalores y los vectores propios son calculados usando la correlación de la matriz M. Los nuevos valores del atributo son luego calculados al multiplicar M con los vectores propios de Corr(M).
- PCA3: Cada variable de M es normalizada por la estandarización de su desviación. Estos valores normalizados son usados para calcular los autovalores y los vectores propios(no hay diferencia entre los coeficientes de covarianza y correlación) y también para el cálculo de los nuevos atributos.

6. PROBLEMA 1: ENCUESTA SOBRE ACCIDENTES EN EMPRESAS MINERAS

Tenemos una tabla que recoge los resultados de una encuesta en varias empresas mineras, sobre las circunstancias que rodearon la ocurrencia de un suceso (variable S) que puede ser Accidente o Incidente en función de su gravedad.

OBJETIVO del análisis: Determinar las condiciones asociadas a uno u otro tipo de suceso con objeto de conocer su casuística y adoptar medidas preventivas en su caso.

Para realizar el análisis tenemos que poner como variable respuesta la variable suceso.

En este problema tenemos las siguientes variables con sus correspondientes etiquetas:

- Variable respuesta: SUCESO (S)
- Variables explicativas: HORA (H), DÍA (D), MES (M), NACIONALIDAD (Na), TIPO DE CONTRATO (TC), TIEMPO EN OBRA (TO), PUESTO DE TRABAJO (PT) FORMACIÓN (F), COMUNIDAD AUTÓNOMA (CA) RÉGIMEN (R), PLAZO DE EJECUCIÓN (PE), DIRECCIÓN Y SUPERVISIÓN (DS)

Hay que tener especial cuidado con las variables COMUNIDAD AUTÓNOMA, RÉGIMEN y DIRECCIÓN Y SUPERVISIÓN pues ,en principio, se duda de que exista suficiente representación para cada tipo de suceso.

6.1. APLICACIÓN: ALGORITMOS DE PONDERACIÓN BINARIA

6.2. APLICACIÓN: ALGORITMOS DE PONDERACIÓN CONTÍNUA

En este apartado comparamos la solución dada por diferentes métodos con todas las variables y usando toda la muestra de entrenamiento. En la presentación(Selección_de_variables.ppt) además de esto tambien se muestran la diferencia que hay entre las soluciones:

- Usando todas las variables y sin las variables CA,R y DS
- Usando validación cruzada y sin usar validación cruzada.

En los casos en que la diferencia sea notable seránmencionados también en el documento.

Método	solución
One Attribute Eval	CA,R,TO,F,PT,H,TC,RP,HO,D,FP,M,DS,Ed,CT,An,ER,Na,E
RelievefF Attribute Eval	CA,H,ER,R,HO,TC,PT,Ed,TO,RP,D,FP,M,An,F,DS,E,Na,CT

6.3. APLICACIÓN: LATENT SEMANTIC ANÁLISIS

La variable encuesta es la variable latente

6.4. APLICACIÓN COMPONENTES PRINCIPALES

Valores propios (eigenvalue)	Proporción varianza explicada(proportion)	Proporción acumulada(cumulative)	Componentes principales
7.19233	0.06365	0.06365	1
5.05384	0.04472	0.10837	2
4.54358	0.04021	0.14858	3
4.27477	0.03783	0.14858	4
3.97975	0.03522	0.18641	5
…	…	….	….
1.01639	0.00899	0.91005	51
1.01639	0.00899	0.91905	52
1.01639	0.00899	0.92804	53
1.01639	0.00899	0.93704	54
1.01639	0.00899	0.94603	55
1.01639	0.00899	0.95503	56

- En este caso, tendríamos que quedarnos con 56componentes principales para poder explicar un 95,5% de la varianza.
- Antes de realizar el análisis teníamos 116 variables. Contando que cada categoría es una nueva variable.

7. PROBLEMA 2: CREDIT-SCORING

Los bancos están interesados en saber si los clientes le van a pagar el crédito o no.

El objetivo de credit-scoring es modelar o predecir la probabilidad de que un cliente con ciertas características esté considerado como un potencial riesgo.

Nuestro conjunto de datos consiste en 1000 personas que tienen un crédito en un banco alemán. Para cada cliente la binanaria variable respuesta "creditability" está disponible . Además, fueron registradas 20 covariables que influyen en la variable respuesta.

- A la hora de tratar con este conjunto de datos tenemos que cambiar variables que están como numéricas y ponerlas como nominales.
- Para ello usamos el filtro no supervisado atributo Numerical to nominal.

  Las variables: Laufteint, Hoehe y Alter son las únicas numéricas

- Descripción de las variables
  1. Laufkont: balance de la cuenta corriente
  2. Laufzeit: duración en meses
  3. moral : pagamiento de créditos previos
  4. Verw: propósito del crédito
  5. hoehe: cantidad de crédito en "Deutsche Mark" (metric)
  6. sparkont: valores de los ahorros
  7. Beszeit: Has estado empleado durante….
  8. rate: plazo en % de ingresos seguros
  9. famges :estado social/sexo
  10. buerge : nuevos deudores/fiadores
  11. Wohnzeit: viviendo en una casa familiar durante…
  12. verm:posesiones
  13. alter : edad en años
  14. weitkred : nuevos créditos rápidos
  15. Wohn: tipo de apartamento
  16. bishkred: número de creditos previos pedidos a este banco(incluidos los créditos rápidos)
  17. beruf : ocupación
  18. pers : número de personas que mantienes
  19. telef : ¿tienes teléfono?

20. Gastarb:¿trabajador extranjero?
21. kredit : buen crédito o no

7.1. APLICACIÓN: ALGORITMOS DE PONDERACIÓN BINARIA

7.2. APLICACIÓN: ALGORITMOS DE PONDERACIÓN CONTÍNUA

En este apartado comparamos la solución dada por diferentes métodos con todas las variables y usando toda la muestra de entrenamiento. En la presentación(Selección_de_variables.ppt) además de esto tambien se muestran la diferencia que hay entre las soluciones

- Usando validación cruzada y sin usar validación cruzada.

En los casos en que la diferencia sea notable serán mencionados también en el documento.

Método	solución
One Attribute Eval	3,2,9,11,10,6,8,7,18,17,20,19,12,13,16,14,15,4,1,5
RelievefF Attribute Eval	1,3,4,9,7,6,12,11,19,8,17,2,16,10,18,5,13,14,15,20

7.3. APLICACIÓN: LATENT SEMANTIC ANÁLISIS

Se reduce la dimensión: pasamos de 21 a 2

1. Laufkont: balance de la cuenta corriente
2. Laufzeit: duración en meses

Son las variables escogidas

7.4. APLICACIÓN COMPONENTES PRINCIPALES

Valores propios(eigenvalue)	Proporción varianza explicada(proportion)	Proporción acumulada(cumulative)	Componentes principales
4.03815	0.05938	0.05938	1
3.30805	0.04865	0.10803	2
2.71972	0.04	0.1480	3
…	….	….	….
0.72783	0.0107	0.93203	45
0.70462	0.01036	0.94239	46
0.67681	0.00995	0.95234	47

- En este caso, tendríamos que quedarnos con 47 componentes principales para poder explicar un 95,23% de la varianza.
- Pasamos de 128 variables a 47 componentes principales.