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

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.

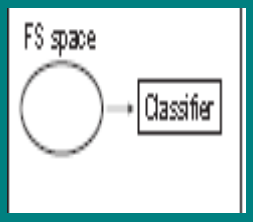
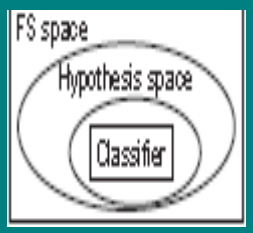
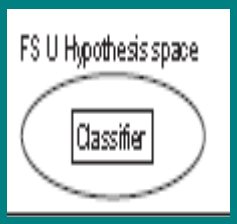
Los principales objetivos del problema de selección de variables son:

- Evitar el sobreajuste y mejorar el modelo obtenido, es decir, la predicción obtenida en caso de clasificación supervisada y mejor selección de variables ("cluster") en caso de que se quieran seleccionar variables.
- Proporcionar modelos más rápidos y con más coste-eficad
- Obtener más profundidad en los procesos subyacentes que en el conjunto de datos generado.

Sin embargo, las ventajas de las técnicas de selección de variables tienen un cierto precio, por ejemplo, la búsqueda de un conjunto de variables relevantes introduce complejidad en la tarea de modelar un problema. En vez de solo optimizar los parámetros de todo el conjunto de datos, ahora encontraremos también los parámetros óptimos del modelo para el subconjunto de variables óptimo, porque no hay garantía de que los parámetros óptimos para todo el conjunto de variables sean igualmente óptimos para el subconjunto de óptimo de variables. Por tanto, la búsqueda del el espacio de hipótesis del modelo aumenta en otra dimensión: encontrar el subconjunto óptimo de variables relevantes. Las técnicas de selección de variables se diferencian unas de otras por la manera en que ellas incorporan esta búsqueda del subjonjunto de variables en el modelo de selección.

En el contexto de clasificación, las técnicas de clasificación de variables pueden ser organizadas en tres categorías, dependiendo de cómo combinan la selección de variables con la construcción de la clasificación del modelo: *métodos filter* , *métodos wraper* y *métodos embedded.*

Modelos de búsqueda	
FILTER	Calculan la relevancia de las variables mirando solo las propiedades intrínsecas de los datos. En la mayoría de los casos, se calcula la relevancia y las variables con menor relevancia son eliminadas.
WRAPPER	Es definido el subconjunto de posibles variables en el espacio de búsqueda de un procedimiento, y varios subconjuntos de variables son generadas y evaluadas. La evaluación de un subconjunto especifico de variables es obtenido intentando y testeando un modelo de clasificación especifico, interpretando esta aproximación especificaremos el algoritmo de clasificación. La búsqueda de todas los subconjuntos de variables, una búsqueda de un algoritmo es un "wrapped" alrededor de la clasificación del modelo. Aunque, como el espacio de subconjuntos de variables crece exponencialmente con el número de variables, los métodos de búsqueda heurística son usados para guiar la búsqueda del subconjunto óptimo.
EMBEDDED	El subconjunto óptimo de variable es obtenido en la construcción del clasificador, y puede ser visto como una búsqueda en la combinación del espacio de los subconjuntos de características y hipótesis.

Modelo de búsqueda			Ventajas	Desventajas	Ejemplos
Filter		Univariante	Rápido Escalable Independiente del clasificador	Ignora las dependencias entre las variables Ignora la interacción con el clasificador	Chi-square Distancia euclidea t-test Information gain Gain ratio
		Multivariante	Modelos de variables dependientes Independiente del clasificador Computacionalmente es mejor que los métodos wapper	Más lento que las técnicas univariantesion Menos escalable que las técnicas univariantes. Ignora la interacción con el clasificador	selección de variables basada en correlación (CFS) Markov blanket filter (MBF) selección de variables basada en correlación rápida (FCBF)
Wrapper		Determinista	Simple Interactúa con el clasificador Modelos de variables dependientes Computacionalmente menos intensivo que los métodos aleatorizados.	Riesgo de sobreajuste Más propenso que los algoritmos aleatoriezados a quedarse atascados en un optimo local. (greedy search) El clasificador depende de la selección	Sequencial forward selection (CFS) Sequencial backward elimination (SBE) Añadir q quitar r Beam search
		Aleatorio	Menos propenso a la estimación local Interacciones con el clasificador Modelos de variables dependientes	Intensivo computacionalmente Clasificador depende de la selección El riesgo de sobreajuste es más elevado que en los algoritmos deterministas	Estimación de distribución de los algoritmos. Algoritmos genéticos
Embedded			Interacciona con el clasificador Computacionalmente es mejor que los métodos wapper Modelos de variables dependientes	El clasificador depende de la selección.	Arboles de decisión Selección de variables usando el peso del vector de SVM

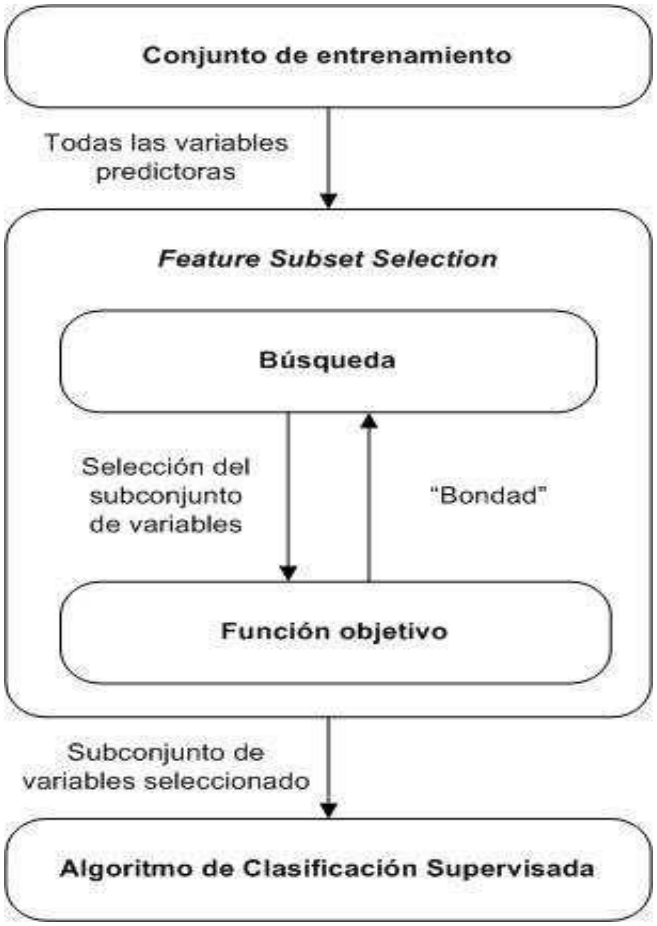
Vamos a estudiar el problema de selección de variables a través de algoritmos de selección de variables. El problema de desarrollar un ASV es básicamente uno de búsqueda en un espacio de estados. Cada estado representa un subconjunto de variables ponderadas; el objetivo es encontrar el estado con la 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).

Esquema del procedimiento de selección de subconjuntos de variables para problemas de clasificación supervisada



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.

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".

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:

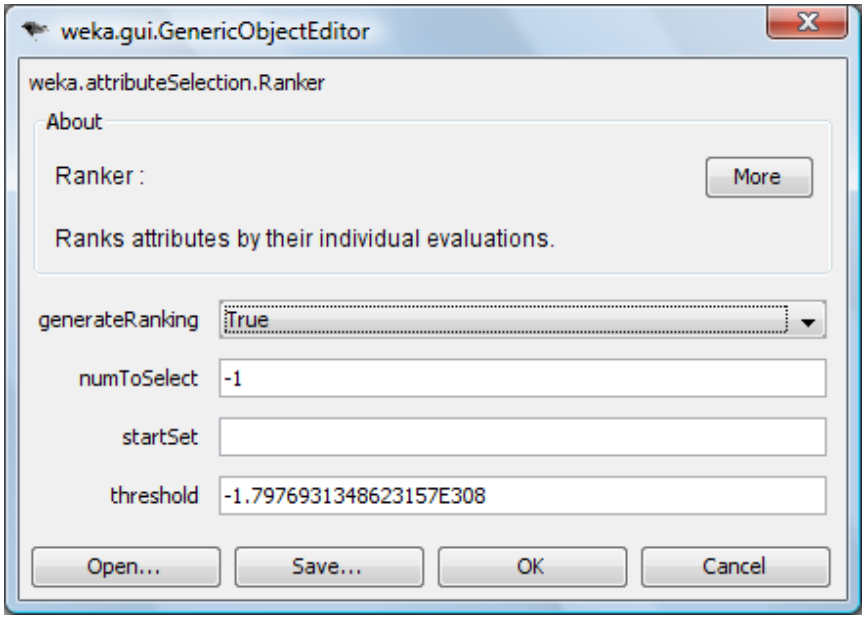
2.7. SYMMETRICAL UNCERT ATTRIBUTE SET EVAL

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 individiduales. 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.



- **Determina el rango de los atributos(variables) por sus evaluaciones individiduales.** Se usa en conjunción con los evaluadores que ordenan los atributos por relevancia. (ReliefF, GainRatio, Entropy etc).
- **OPTIONS**
  - **generateRanking** – Es una constante opcional. Ranker es solo capaz de generar atributos ranking
  - **numToSelect** – Especifica el número de atributos a retener. Por defecto viene el valor -1 que indica que todos los atributos son retenidos. Usa otra opción o un umbral para reducir la colección de atributos.
  - **startSet** – Especifica una colección de atributos a ignorar. Cuando generamos el ranking, Ranker no evalúa los atributos en esta lista . Esto se especifica con una lista de atributos separada por comas, empezando en 1. Se puede incluir intervalos. Ejemplo. 1,2,5-9,17.
  - **Threshold –** Se fija el umbral por el cúal se pueden descartar atributos. El valor que viene dado por defecto no descarta ningún atributo. Se usa esta opción o numToSelect para reducir la colección de atributos

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.

3.7.1. En Weka

- Evalúa el valor de una variable usando el clasificador OneR.
- OPCIONES:
  - **evalUsingTrainingData** -- Utiliza los datos de entrenamiento para evaluar atributos mejores que validación cruzada.
  - **folds** -- Número de pliegues para validación cruzada.
  - **minimumBucketSize** -- El número mínimo de objetos en una clase (pasado a OneR).
  - **seed** --Semilla que usamos en validación cruzada.

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 $\theta(pmn)$ porque calcula la distancia entre X y cada uno de los n datos, tomando $\theta(p)$ veces, para determinar su Near-miss and Near-hit dentro de un bucle iterativo m veces. m es una constante que afecta a la exactitud de los niveles de relevancia. Luego, m es escogido independientemente de p y n, la complejidad está en $\theta(pn)$. 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.

El pseudocódigo del algoritmo es el siguiente

Relief(S, m,ζ)

Separamos S en dos S^+={datos positivos} y S^-={datos negativos}

W=(0,0,…..,0)

Desde i =1 hasta m

Se escoge aleatoriamente XϵS

Se escoge aleatoriamente un dato positivo próximo a X, $Z^+\epsilon S^+$

Se escoge aleatoriamente un dato negativo próximo a X, $Z^-\epsilon S^-$

Si (X es positivo)

luego Near-hit= Z^+; Near-miss= Z^-

sino Near-hit= Z^-; Near-miss= Z^+

update-weight(W, X, Near-hit, Near-miss)

Relevancia=(1/m)W

Desde i=1 hasta p

si(relevancia$_i \geq \zeta$)

luego f_i es una variable relevante

sino f_i no es una variable relevante

update-weigth(W, X, Near-hit, Near-miss)

desde i=1 hasta p

$W_i=W_i - diff(x_i, near\text{-}hit_i)^2 + diff(x_i, near\text{-}miss_i)^2$

3.8.1. En Weka

- 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.
- RELIEF:describe un algoritmo estadístico de selección de características que usa muestras de entrenamiento para asignar peso relevante a cada característica.
- OPCIONES:
    - **numNeighbours** – Número de vecinos más cercanos para los atributos estimados.
    - **sampleSize** – Número de casos de muestra. Por defecto(-1) indica que todos los casos serán utilizados para la estimación de los atributos.
    - **seed –**Semillas aleatorias para el muestreo de casos.
    - **sigma** –Conjunto de influencia de vecinos más cercanos. Utiliza una función exponencial para controlar la rapidez de disminución del peso de los casos más distantes. Uso junto con weightByDistance. Valores aconsejados= 1/5 a 1/10 el númeso de vecinos más cercanos
    - **weightByDistance** –proporciona el peso para los vecinos más cercanos

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 parecidos.

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.

4.1. EN WEKA

- Se realiza el “latent semantic analysis” y la transformación de los datos.
- Se usa junto una búsqueda Ranker. Un bajo rango aproximado de todo el conjunto de datos es encontrado por la dspecificación de valores singulares
- OPCIONES:
    - **maximumAttributeNames** –El máximo número de atributos a incluir en la transformación de los nombres de atributos .
    - **normalize** –Normaliza los datos.
    - **rank** – Rango de la matriz que se usa para la reducción de los datos. Puede ser una proporción indicada para la cobertura deseada.

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.

5.1. EN WEKA

- Realiza un análisis de componentes principal y la transformación de los datos. Se emplea en conjunción con una búsqueda de Ranker. La reducción de dimensionalidad se logra escogiendo bastantes vectores propios para considerar para algún porcentaje de la discrepancia en los datos originales---la falta 0.95 (el 95 %). El ruido de atributo puede ser filtrado por la transformación el espacio de la componente pricipal, eliminando algunos de los peores vectores propios, y luego devolviendolos al ámbito original.
- Considera cada nivel de la variable como una variable, es decir, si tenemos una variable sexo con dos niveles: hombre y mujer. Pues al aplicar componentes principales consideramos que sexo_hombre es una variable y sexo_mujer sería otra variable.

- OPTIONS
    - **maximumAttributeNames** –El máximo número de atributos a incluir en la transformación de los nombres de atributos .
    - **normalize** – Normalizar los datos
    - **transformBackToOriginal** – Transformación del espacio de datos y devolviendolo al ámbito original. Si sólo son retenidas las n mejores componentes principales (poniendo varianceCovered < 1) entonces esta opción dará una colección en el espacio original, pero con menos ruido de atributo.
    - **varianceCovered** --Conserve bastantes componentes principales de los atribuutos para considerar para esta proporción de discrepancia.

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. además de esto también 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.

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

6.2.1. Comparación de resultados eliminando y no eliminando las variables CA, R y DS

6.2.1.1. Chi Square Attribute Eval

6.2.1.2. Cost Sensitive Attribute Eval

6.2.1.3. Filtered Attribute Eval

6.2.1.4. Gain Ratio Attribute Eval

6.2.1.5. Info Gain Attribute Eval

6.2.1.6. OneR Atttribute Eval

- Selección de atributos no eliminando las variables CA,R y DS(15,16,19):

  15,16,10,12,11,2,9,13,5,3,14,4,19,6,18,8,17,7,1 : 19
- Selección de atributos eliminando las variables CA, R y DS:

  12,10,11,2,9,13,5,14,3,4,6,8,18,17,7,1 : 16.

Observamos que las variables siguen en el mismo orden excepto las variables señaladas en azul.

6.2.1.7. RelievefF Attribute Eval

- Selección de variables

  15,2,17,16,5,9,11,6,10,13,3,14,4,8,12,19,1,7,18 : 19
- Selección de variables:

  2,17,5,9,11,6,10,13,3,14,4,8,12,1,7,18 : 16

Las variables mantienen el orden de relevancia si se eliminan las variables Ca,R y DS

6.2.1.8. Symmetrical Uncert Attribute Eval

6.2.2. Comparación de resultados con y sin validación cruzada

6.2.2.1. Chi Square Attribute Eval

6.2.2.2. Cost Sensitive Attribute Eval

6.2.2.3. Filtered Attribute Eval

6.2.2.4. Gain Ratio Attribute Eval

6.2.2.5. Info Gain Attribute Eval

6.2.2.6. OneR Atttribute Eval

- Selección de atributos:
  15,16,10,12,11,2,9,13,5,3,14,4,19,6,18,8,17,7,1
- Selección de atributos:

average merit	average Rank	Atributos
87.097 +- 1.336	1 +- 0	15 CA
77.416 +- 1.48	2.1 +- 0.3	16 R
74.192 +- 0.917	4.2 +- 1.08	12 F
73.117 +- 1.609	5.3 +- 2.05	10 TO
72.737 +- 3.304	5.8 +- 3.52	11 PT
70.964 +- 4.154	8.4 +- 3.8	5 HO
70.968 +- 0.692	8.9 +- 1.7	13 RP
70.071 +- 4.011	9.2 +- 4.62	14 FP
69.89 +- 1.392	9.4 +- 2.76	9 TC
69.536 +- 2.228	10.3 +- 2.24	2 H
68.994 +- 1.999	11 +- 2.37	19 DS
67.558 +- 4.52	11.5 +- 4.1	17 ER
67.214 +- 3.001	12.2 +- 4.09	3D
66.675 +- 2.687	13.2 +- 3.37	18 CT
65.769 +- 3.751	14.1 +- 3.18	6 Ed
65.945 +- 3.438	14.4 +- 2.8	8 An
64.153 +- 3.952	14.6 +- 3.88	7 Na
63.792 +- 3.705	15.4 +- 3.14	4 M
29.032 +- 0.692	19 +- 0	1 E

Se mantiene en gran medida el orden de relevancia de las variables tanto con cómo sin validación cruzada. Las que eran muy relevantes sin validación cruzada con validación cruzada lo siguen siendo y las que no eran muy relevantes sin validación cruzada, ahora tampoco lo son. Aunque hay excepciones, por ejemplo las variables 3 y 4(MES (M), NACIONALIDAD (Na)) que son muy relevantes sin validación cruzada y poco relevantes sin validación.

6.2.2.7. RelievefF Attribute Eval

- Selección de variables sin validación cruzada:

  15,2,17,16,5,9,11,6,10,13,3,14,4,8,12,19,1,7,18 : 19

- Selección de variables

average merit	average rank	attribute
0.316 +- 0.039	1 +- 0	15 CA
0.135 +- 0.025	3.3 +- 0.9	2 H
0.129 +- 0.021	3.5 +- 1.2	17 ER
0.121 +- 0.024	4.1 +- 1.87	5 HO
0.124 +- 0.026	4.1 +- 1.81	16 R
0.104 +- 0.013	5.9 +- 1.04	11 PT
0.09 +- 0.014	7 +- 1.1	9 TC
0.077 +- 0.01	7.9 +- 1.14	6 Ed
0.062 +- 0.017	9.5 +- 1.8	10 TO
0.057 +- 0.018	10.5 +- 2.38	13 RP
0.048 +- 0.025	11.2 +- 2.89	14 FP
0.037 +- 0.012	12.6 +- 1.62	12 F
0.034 +- 0.009	13.1 +- 1.14	3 D
0.035 +- 0.013	13.1 +- 1.58	4 M
0.028 +- 0.004	14 +- 0.77	8 An
0.005 +- 0.009	16.5 +- 0.92	19 DS
0 +- 0	17.2 +- 0.6	1 E
0.005 +- 0.017	17.5 +- 1.2	7 Na
0.011 +- 0.018	18 +- 2.05	18 CT

Se mantiene en gran medida el orden de relevancia de las variables tanto con cómo sin validación cruzada. Las que eran muy relevantes sin validación cruzada con validación cruzada lo siguen siendo y las que no eran muy relevantes sin validación cruzada, ahora tampoco lo son.

6.2.2.8. Symmetrical Uncert Attribute Eval

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

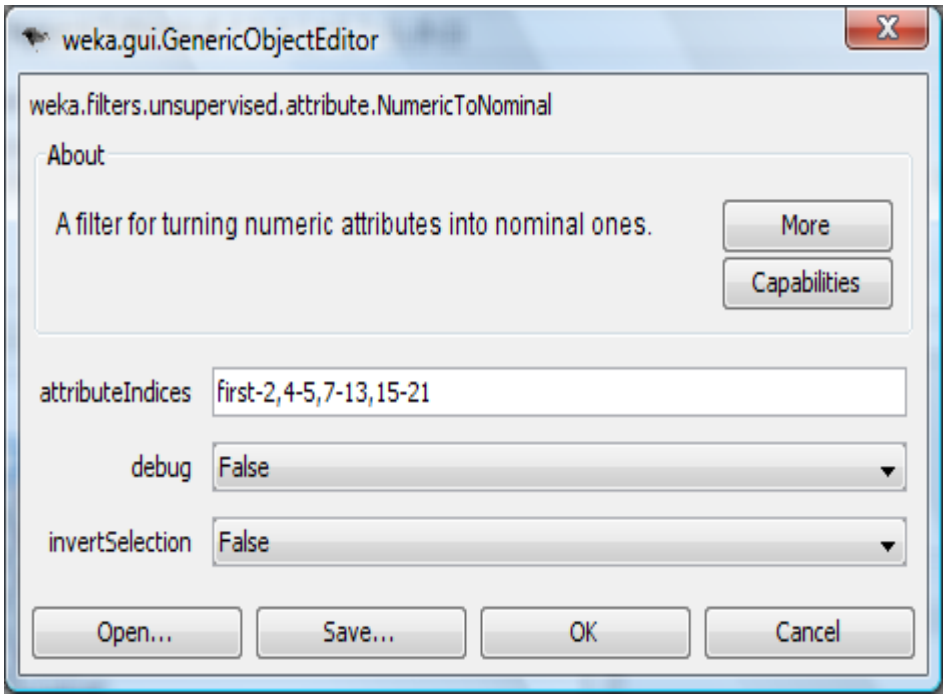
Los bancos están interesados en saber si los clientes le van a pagar el crédito o no.

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: Laufzeit, 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. Además de esto también 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

7.2.1. Comparación de resultados con y sin validación cruzada

7.2.1.1. Chi Square Attribute Eval

7.2.1.2. Cost Sensitive Attribute Eval

7.2.1.3. Filtered Attribute Eval

7.2.1.4. Gain Ratio Attribute Eval

7.2.1.5. Info Gain Attribute Eval

7.2.1.6. OneR Atttribute Eval

- Selección de variables:

  3,2,9,11,10,6,8,7,18,17,20,19,12,13,16,14,15,4,1,5 : 20

- Selección de variables:

average merit	average rank	attribute
71.633 +- 0.258	1 +- 0	3 moral
70 +- 0	4.1 +- 0.3	11 wohnzeit
70.511 +- 0.495	4.3 +- 5.02	2 laufzeit
69.911 +- 0.267	5.6 +- 4.54	9 famges
70 +- 0	5.9 +- 0.3	6 sparkont
70 +- 0	7.1 +- 1.45	7 beszeit
70 +- 0	7.3 +- 0.64	8 rate
70 +- 0	8.6 +- 5.41	12 verm
70 +- 0	9.3 +- 0.64	18 pers
69.811 +- 0.233	9.5 +- 6.64	10 buerge
70 +- 0	9.9 +- 0.7	17 beruf
70 +- 0	11.3 +- 0.64	20 gastarb
70 +- 0	11.9 +- 0.7	19 telef
70 +- 0	13.2 +- 0.75	14 weitkred
70 +- 0	15 +- 0.89	15 wohn
69.122 +- 0.658	15.7 +- 5.08	1 laufkont
69.778 +- 0.131	16.1 +- 0.83	16 bishkred
69.511 +- 0.291	16.9 +- 1.64	13 alter
69.256 +- 0.636	17.3 +- 1.35	4 verw
66.122 +- 0.817	20 +- 0	5 hoehe

Se mantiene en gran medida el orden de relevancia de las variables tanto con cómo sin validación cruzada. Las que eran muy relevantes sin validación cruzada con validación cruzada lo siguen siendo y las que no eran muy relevantes sin validación cruzada, ahora tampoco lo son. Aunque hay excepciones como la variable 10 (buerge : nuevos deudores/fiadores) muy relevante sin validación cruzada y poco relevante con validación cruzada.

7.2.1.7. RelievefF Attribute Eval

- Selección de atributos:

  1,3,4,9,7,6,12,11,19,8,17,2,16,10,18,5,13,14,15,20 : 20

- Selección de variables

average merit	average rank	attribute
0.153 +- 0.011	1 +- 0	1 laufkont
0.066 +- 0.005	2 +- 0	3 moral
0.048 +- 0.004	3.1 +- 0.3	4 verw
0.039 +- 0.005	4.4 +- 0.8	9 famges
0.036 +- 0.004	5.4 +- 0.92	7 beszeit
0.032 +- 0.004	6 +- 1.18	6 sparkont
0.029 +- 0.005	6.7 +- 1.42	12 verm
0.024 +- 0.004	8.8 +- 2.09	11 wohnzeit
0.022 +- 0.005	9.4 +- 1.74	8 rate
0.019 +- 0.003	10.8 +- 1.66	19 telef
0.018 +- 0.004	11.4 +- 2.33	17 beruf
0.018 +- 0.002	11.5 +- 1.75	2 laufzeit
0.016 +- 0.002	12.5 +- 1.57	16 bishkred
0.016 +- 0.003	12.7 +- 2	10 buerge
0.012 +- 0.003	15.6 +- 1.85	18 pers
0.011 +- 0.001	16.2 +- 0.6	5 hoehe
0.01 +- 0.001	17 +- 0.89	13 alter
0.009 +- 0.003	17.5 +- 1.57	15 wohn
0.008 +- 0.003	18.2 +- 1.47	14 weitkred
0.003 +- 0.001	19.8 +- 0.4	20 gastarb

Se mantiene en gran medida el orden de relevancia de las variables tanto con cómo sin validación cruzada. Las que eran muy relevantes sin validación cruzada con validación cruzada lo siguen siendo y las que no eran muy relevantes sin validación cruzada, ahora tampoco lo son.

7.2.1.8. Symmetrical Uncert Attribute Eval

7.3. APLICACIÓN: LATENT SEMANTIC ANÁLISIS

Se reduce la dimensión: pasamos de 21 a 2

Laufkont: balance de la cuenta corriente

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.

8. APLICACIONES EN BIOINFORMÁTICA

8.1. SELECCIÓN DE VARIABLES PARA UN ANÁLISIS DE SECUENCIAS

El análisis de secuencias tiene una larga tradición en bioinformática. En este contexto de selección de variables, se pueden distinguir dos tipos de problemas:

- **Content analysis** (análisis de contenidos), está enfocado hacia las caracterísricas de una secuencia, codifica la tendencia de las proteínas o satisfacción de una cierta función biológica.
- **Signal analysis** (análisis de señales), se centra en la identificación de importantes elementos en la secuencia, como son los elementos estructurales de un gen o elementos regulares.

8.1.1. Content analysis

La predicción de subsecuencias que codifican las proteínas ha sido un foco de interés desde el inicio de la bioinformática. Puesto que, muchas características pueden ser extraídas de una secuencia, y la dependencia más importante ocurre entre posiciones adyacentes, para ello se desarrollaron muchas variaciones en los modelos de Markov. Tratar con grandes cantidades de posibles características, introdujo el modelo interpolado de Markov (INM), el cual usó la interpolación entre diferentes ordenes del modelo de Markov con muestras de pequeño tamaño, y un método filtro("filter method") que selecciona solo las características relevantes.

Además, se extendió el INM a la situación de tratar con las dependencias de las características no adyacentes, resultando el modelo contextualmente interpolado (ICM), el cual mezcla un árbol de decisión Bayesiano con un método filter (Chi-square) para obtener las características relevantes. Recientemente, el camino de las técnicas de selección de variables(FS) para codificar la predicción potencial fueron obtenidas combinando diferentes medidas del código de la predicción potencial, y luego se usó la aproximación del filtro multivariante "blanket" de Markov (MBF) para retener solo las relevantes.

Una segunda clase de técnicas buscan la función predicción de las proteínas de la secuencia. Esto combina un algoritmo genético con un test Gamma para obtener el subconjunto de variables para la clasificación de grandes cantidades de subunidades de rRNA, inspirado en la rebúsqueda de de técnicas FS para conseguir importantes subconjuntos de amino ácidos que describan la clase funcional de la proteína. Una interesante técnica usa el núcleo escalado para la SVM como un método que calcula el peso de las variables, y remueve las subsecuencias de variables con menor peso.

El uso de las técnicas FS en el ámbito del análisis de secuencias es introducido en numerosas aplicaciones recientes.

8.1.2. Signal analysis

Muchas metodologías de análisis de secuencias que reconocen más o las señales conservadas en la secuencia, representando sitios para varias proteínas o proteínas complejas. Una aproximación para encontrar motivos de reglamentación, es relacionar los motivos a los niveles de expresiones genéticas usando una regresión. La selección de variables puede ser usada para buscar los motivos que maximizan el ajuste de un modelo de regresión. Una clasificación aproximada es escogida para discriminar motivos. El método usa el umbral de números de errores de clasificación para valorar los genes mediante la relevancia en la clasificación. Del valor TNoM, se obtiene un p-valor que representa la significación de cada modelo. Los motivos son clasificados de acuerdo con su p-valor.

Otra línea de búsqueda es transformar en el contexto del conjunto de genes predicho, donde estructurales elementos como tales se trasladaron del sitio inicial (TIS) y los sitios de empalme son modelados como problemas específicos de clasificación. El problema de selección de variables para el reconocimiento de elementos estructurales. El problema de predicción de los sitios de empalme se resuelve combinando el método backward junto con el criterio de evaluación SVM para valorar la relevancia de las variables. Una estimación de la distribución del algoritmo fue usada para ganar más perspicacia de las relevantes características en la predicción de los sitios de embarque.

En el futuro, se espera que las técnicas de selección de variables mejoren las técnicas de predicción, las cuales identifican las relevantes variables relacionadas con los sitios de empalme y la alternativa TIS.

8.2. SELECCIÓN DE VARIABLES APLICADO AL ANÁLISIS DE MICROARRAY

Durante la última década, la aparición de conjuntos de datos de microarrays ha estimulado una nueva línea de investigación en bioinformática. El conjunto de datos de microarrays(“microarrays data”) constituye un gran desafío para las técnicas de cálculo, debido a sus grandes dimensionalidad (hasta varias decenas de miles de genes) y a su pequeño tamaño de muestra. Además, las complicaciones experimentales, como el ruido y la variabilidad hacen del análisis del conjunto de datos de microarrays un dominio emocionante.

Con el fin de hacer frente a estas características particulares del análisis del conjunto de datos de microarrays, se necesitaba emplear técnicas de reducción de la dimensión y pronto su aplicación se convirtió en una de estándar de facto en el campo. Para ello se han construido nuevas metodologías y adaptado las conocidas FS.

8.2.1. El paradigma del filtro univariado: simple pero eficiente

Debido a la alta dimensionalidad de la mayoría de los análisis de microarrays, las técnicas que han captado la mayor interés son las técnicas FS por su eficiencia y rápidez como los métodos de filtro univariantes. La prevalencia de estas técnicas univariantes ha dominado el campo, y hasta ahora, se comparan las evaluaciones de diferentes clasificaciones y técnicas FS sobre el conjuntos de datos de microarrays de ADN centrados únicamente en el caso univariante. Este interés por la aproximación univariante se puede explicar por varias razones:

• La salida proporcionada por el ranking univariante de características es intuitiva y fácil de entender;

• el gen de la salida de clasificación podría cumplir los objetivos y el las expectativas de los expertos en el campo biológico quienes esperan tener el resultado de la secuencia validada mediante técnicas de laboratorio o explorar búsquedas bibliográficas. Los expertos no tienen la necesidad de emplear técnicas de selección que tengan en cuenta gen interacciones;

• la falta de conocimiento posible de los subgrupos de expertos en el campo de la expresión genética de la existencia de técnicas de análisis de datos multivariantes;

• el tiempo de cálculo extra necesario para las técnicas de selección genéticas multivariantes.

Se ha desarrollado una amplia gama de características nuevas o adaptadas univariado técnicas de clasificación desde entonces se ha desarrollado. Estas técnicas se pueden dividir en dos clases: modelos paramétricos y los métodos de modelado libre.

Los métodos paramétricos asumen una determinada distribución a partir de la cual se han generado muestras(observaciones). El t-test y ANOVA se encuentran entre las técnicas más utilizadasen los estudios de microarrays, aunque se usan de forma básica, posiblemente sin justificación de sus principales hipótesis, esto no esadmisible. Las modificaciones del t-test estándar para poder atender mejor con muestras de pequeño tamaño y el ruido inherente de conjuntos de datos de expresión genética incluye una serie de T o t-test como las estadísticos (se diferencian principalmente en la forma en que se estima la varianza) y una serie de Marcos Bayesiano . Aunque las hipótesis Gausianas handominado el campo, otros tipos de enfoques paramétricos pueden también ser encontrados en la literatura, tales como modelos de regresión enfoquesy modelos de distribución Gamma.

Debido a la incertidumbre acerca la distribución real subyacente de escenarios de expresión de muchos genes, y las dificultades para validar las suposiciones de distribución, debido al pequeño tamaño de muestral, no paramétrica o modelo de los métodos de libre han sido ampliamente propuestos como una alternativa atractiva para hacer menos estrictos supuestos de distribución . Muchos de los parámetros del modelo libre, frecuentemente prestados del campo de la estadística, han demostrado su utilidad en muchos estudios de expresión génica, incluyendo la prueba de la suma de los rangos de Wilcoxon, la suma de los cuadrados entre-dentro de las clases(BSS / WSS) y el método de productos de los rangos.

Una clase específica de métodos de modelo libre estima la distribución de referencia del estadístico usando permutaciones aleatorias de los datos, permitiendo el cálculo de un modelo de versión libre asociado a pruebas no paramétricas. Estas técnicas han surgido como una sólida alternativa para hacer frente a las especificidades de los datos de microarrays de ADN, y de no dependen de fuertes supuestos paramétricos. Su principio de permutación, en parte alivia el problema de las muestras de pequeño tamaño en los estudios de microarrays, la mejora de la robustez frente los valores extremos.

También menciona prometedoras métricas de tipo no-paramétrica las cuales, en lugar de tratar de identificar los genes expresados diferencialmente en el nivel de la población en su conjunto (por ejemplo, la comparación de medias de la muestra), son capaces de para capturar los genes que son significativamente más desregulados en sólo un subconjunto de muestras. Estos tipos de métodos ofrecen una aproximación más específica para la identificación de marcadores, y puede seleccionar genes que presentan patrones complejos. Además, también señalan la importancia de los procedimientos de para el control de los diferentes tipos de errores que se presentan en este complejo de escenario de múltiples pruebas de miles de gen, con un enfoque especial sobre las contribuciones para el control de la falsa tasa de descubrimiento (FDR).

8.2.2. Hacia modelos más avanzados: el paradigma multivariado para el filtro(filter), técnicas de envoltura e incrustados(wrapper, embedded)

Los métodos de selección univariantes tienen ciertas restricciones y pueden conducir a clasificadores menos precisos, no tienen en cuenta la interacción entre genes. Así, los investigadores han propuesto técnicas que tratan de capturar estas correlaciones entre los genes.

La aplicación de los métodos de filtro multivariante va de simples interacciones de dos variables, hacia soluciones más avanzadas resultado de explorar las interacciones de orden superior, tales como la correlación en base a características selección (CFS) y diversas variantes de la Markov método de filtro de manta (Markov blanquet filter method).

El procedimiento de selección de variables usando **métodos *wrapper*** (métodos de embase)o **métodos *embedded*** (métodos de incrustado) ofrecen una manera alternativa de realizar una selección múltiple subconjunto de genes, incorporando sesgo del clasificador en la búsqueda y ofreciendo así una oportunidad de construir clasificadores más precisos. En el contexto de el análisis de microarrays, la mayoría de los *métodos wrapper* están basados en búsquedas heurísticas de aleatorias, aunque en algunos casos también se usan técnicas de búsqueda secuencial. Es interesante la aproximación *filtro-wrapper* que es considerada como un híbrido ya que cruza una clasificación de genes pre-ordenados univarantemente con *método wrapper* que incrementa gradualmente.

Otra característica de cualquier procedimiento *wrapper* concierna a la función usada para evaluar cada subconjunto de genes encontrados. Esta medida de evaluación es usada para realizar comparaciones con trabajos previos. Sin embargo, los últimos propuestas que abogan por el uso de métodos para la aproximación de las área bajo la curva ROC, o la optimización de modelo LASSO (Contracción menos absoluto y selección de operador). Curvas ROC proporcionan una medida de evaluación interesante, especialmente adecuada a la demanda para la detección de los diferentes tipos de errores en muchos escenarios biomédica.

Los *métodos embedded* tienen la capacidad de usar varios clasificadores para descartar características y por lo tanto proponer un subconjunto de genes discriminativo. Los *métodos embedded* también utilizan el peso de cada característica en los clasificadores lineales como SVMS y de regresión logística. Estos pesos se utilizan para reflejar la importancia de la cada gen de una manera multivariante, y permitir así la eliminación de genes con poco peso.

En parte debido a la gran complejidad computacional de los *métodos wrapper* y en menor grado a las aproximaciones *embedded*, estas técnicas no han recibido tanto interés como las propuestas de filtro. Sin embargo, en la práctica es útil comprobar si es posible reducir el espacio de búsqueda utilizando un método de filtro univariante, y sólo entonces se aplican los *métodos wrapper o embedded,* ajustándonos al tiempo de computación de los recursosdisponibles.

8.3. RELACIÓN CON LOS ÁMBITOS PEQUEÑA MUESTRA

Tamaños de muestra pequeños, y su riesgo inherente de la imprecisión y el sobreajuste, plantean un gran desafío para muchos problemas de modelización en bioinformática. En el contexto de la selección de características, dos iniciativas han surgido en respuesta a esta experimental situación: el uso de criterios de evaluación adecuados, y el uso de los modelos de selección de características robustos y estables.

8.3.1. Criterios de evaluación adecuados

Varios trabajos han advertido sobre el gran número de aplicaciones que no llevan a cabo una independiente y honesta validación de la exactitud de los porcentajes reportados. Pues no se realiza un proceso de selección de características externas en el entrenamiento de la regla de clasificación en cada etapa que estimamos la precisión de la estimación.

Además, nuevos métodos de estimación de la exactitud de predicción con características prometedoras, como la estimación de errores reforzado, han surgido para hacer frente a las especificidades de los dominios de la muestra pequeño.

8.3.2. Aproximación de emsemble selección de características

En lugar de optar por un método particular, FS, y se aceptar su resultado como el subconjunto final, pueden ser combinados diferentes métodos de FS usando la aproximación emsemble FS. Basado en que a menudo no existe una única técnica universalmente óptima de la selección de características, y debido a la posible existencia de más de un subconjunto de de las características que discrimina a los datos igual de bien, el modelo de combinación de enfoques ha sido adaptado para mejorar la robustez y estabilidad del resultado.

Las nuevas técnicas emsemble en el microarray incluyen un promedio de varios subconjuntos de características destacadas, integrando una colección de expresiones genéticas diferenciales univariantes, utilizando diferentes iteraciones de un algoritmo genético para evaluar la importancia relativa a cada característica, calculando el test de Kolmogorov -Smirnov en diferentes muestras bootstrap para asignar una probabilidad debe ser seleccionado, y un número de aproximaciones Bayesianas de la media. Además, los métodos basados en una colección de árboles de decisión pueden ser utilizado como una emsemble FS para evaluar la relevancia de cada característica.

Aunque el uso de aproximaciones emsemble requiera adicionales recursos computacionales, nos gustaría señalar que ofrecen un marco oportuno para hacer frente a los dominios de muestra pequeña ,proporcionando recursos adicionales de computación que son asequibles.

8.4. SELECCIÓN DE CARACTERÍSTICAS EN LAS PRÓXIMAS DOMINIOS

8.4.1. polimorfismo de nucleótido único análisis

Polimorfismos de nucleótido único (SNP) son mutaciones en una única posición de nucleótidos que se produjo durante la evolución y se aprobaron a través de la herencia, que representan la mayoría de la variación genética entre los diferentes individuos. SNP están en el primer plano de muchos estudios sobre asociaciones entre enfermedad-gen, cuyo número se estima sobre de 7 millones en el genoma humano. Por lo tanto, la selección de un subconjunto de SNP que sea lo suficientemente informativo, pero aún lo suficientemente pequeño para reducir la sobrecarga del genotipo es un paso importante hacia la la asociación enfermedad-gen. Normalmente, el número de SNP considerado no es superior a decenas de miles, con muestras de tamaño cien.

Varios métodos computacionales para la selección htSNP(holotipo SNP; un conjunto de SNPs se encuentra en un cromosoma) se han propuesto en los últimos años. Una aproximación se basa en la hipótesis de que el genoma humano puede ser visto como una serie de bloques discretos que comparten sólo un conjunto muy pequeño de los holotipos más comunes. Esta aproximación apunta a identificar un subconjunto de SNPs que permita distinguir todos los haplotipos más comunes, o al menos explicar un cierto porcentaje de ellos. Otra aproximación de la selección htSNP se basa en las asociaciones de pares de SNPs, y trata de seleccionar un conjunto de htSNPs de tal manera que cada uno de los SNPs de un haplotipo es altamente asociado con uno de los htSNPs . Una tercera aproximación considera a htSNPs como un subconjunto de todos los SNPs, a partir del cual pueden ser reconstruídos los restantes SNP. La idea es seleccionar htSNPs en base de lo bien que predigan el conjunto restante formado por los SNPs no seleccionados.

Una aproximación ensemble es aplicada con éxito a la la identificación de SNPs para el alcoholismo, mientras que proponen una sólida técnica de la selección de características basadas en un híbrido entre un algoritmo genético y la SVM. El algoritmo de selección de variables Relief-F, en relación con tres algoritmos de clasificación (K - NN, SVM y naive Bayes) ha sido propuesto. Algoritmos Genéticos han sido aplicados a la búsqueda del mejor subconjunto de SNPs, evaluandolas con un filtro de múltiples variables (CFS), y también de forma wrapper (con un árbol de decisión). La regresión lineal múltiple SNP (algoritmo de predicción) predice un genotipo completo basado en los valores de las SNPs informativas, sus posiciones entre todos los SNPs, y una muestra de genotipos completa.

8.5. CONCLUSIONES Y PERSPECTIVAS FUTURAS

Los principales problemas que aparecen en el campo de la bioinformática son: la gran dimensionalidad, y los tamaños de muestra pequeños.Para hacer frente a estos problemas, una gran cantidad de técnicas de FS ha sido diseñado por los investigadores en bioinformática, el aprendizaje de máquinas y minería de datos.

Un esfuerzo amplio y fructífero se ha realizado durante los últimos años en la adaptación y la propuesta de técnicas FS de filtro univariante. En general, se observa que muchos investigadores en el campo todavía piensan que aproximaciones FS de filtro se limitan sólo a enfoques univariantes. La propuesta de algoritmos de selección múltivariantes puede ser considerada como una de las líneas de futuro prometedor de trabajo para la bioinformática de la comunidad.

Otra línea de investigación se basa en mejorar la robusted de la solución determinada por la aproximación ensemble. A fin de aliviarlos pequeños tamaños de muestra real de la mayoría de las aplicaciones en bioinformáticas, el desarrollo de estas técnicas, combinadas con los criterios de evaluación adecuados, constituye una interesantedirección para futuras investigaciones FS.

Otras oportunidades interesantes para la investigación futura FS será la extensión hacia ámbitos como la bioinformática próximo SNP, y la combinación de heterogéneas fuentes de datos. Actualmente la selección de variables no son esenciales en este campo aunque se cree que que su aplicación lo será para hacer frente a la alta dimensional de estas aplicaciones.

ANEXO I: PAQUETE 'WilcoxCV'

INTRODUCCIÓN

Hace pocos años, numerosos métodos han sido propuestos basados en microarrays para predecir clases. Aunque muchos de ellos han sido especialmente en el caso de que n<<p (muchas más variables que datos), anteriormente la selección de variables era casi siempre necesaria cuando el número de genes alcanza decenas de miles, es usual es recientes conjuntos de datos. El estadístico de la suma de los rangos de Wilcoxon es, junto con el t-estadístico, una de las estándar aproximaciones para la selección de variables. Es bien conocido que el paso de la selección de variables debe ser visto como una parte de la construcción del clasificador y , el cual, debe ser obtenido basándonos solamente en los datos de entrenamiento.

Cuando la exactitud del clasificador es evaluada vía validación cruzada o validación cruzada con Monte-Carlo, esto significa que tenemos que realizar p Wilcoxon o t-test para cada iteración, la cual comienza a ser una desalentadora tarea debido al incremento de p.

Como consecuencia de ello, muchos autores a menudo realizan la selección de variables usando sólo una vez con todos los datos disponibles, que pueden inducir una dramática subestimación de la tasa de error y así producir informes donde se pierda poder predictivo .Se propuso un método rápido de selección de variables basado en el test de Wilcoxon basado en la validación cruzada y de Monte Carlo de validación cruzada (también lo conocemos como división de azar en el aprendizaje y equipos de prueba). Esta implementación se basa en una simple fórmula matemática utilizando sólo el rango calculado del conjunto de datos original. ("new")

Test de la suma de los rangos de Wilcoxon

La idea de usar un test basado en rangos surge de aplicar la hipótesis de simetría a los test de signos.

La teoría de los tests basados en rangos es más complicada que la del test del signo.
Bajo Ho, el estadístico de un test de rangos puede ser representado como una suma de v.a. independientes pero no idénticamente distribuidas y, bajo H1, se pierde inclusive la independencia. Por ello, necesitaremos nuevas versiones del teorema central del límite.
Se desea realizar el siguiente test

$$H_0: \theta = 0 \qquad \text{vs} \qquad H_1: \theta > 0$$

siendo θ el centro de simetría de X (será además la media si ésta existe).
Supondremos que X1,...,Xn es una muestra aleatoria de una distribución F(x-θ) con Fє Ωs, siendo

Ωs = {F / F es absolutamente continua con única mediana en 0 y simétrica}

El test del signo se basa en información sobre el signo de las observaciones y no utiliza información sobre la distancia de las observaciones al cero. Sin embargo, si la distribución es simétrica alrededor de 0, el vector de valores absolutos $|X_1|, |X_2|,...,|X_n|$ es un estadístico suficiente y por lo tanto, parece razonable tratar de incorporar esta información.

Sea $|X|^{(1)} \leq |X|^{(2)} .... \leq |X|^{(n)}$, la muestra de valores absolutos ordenados y

R_j = rango ($|X_j|$) es decir $|X_j| = |X|^{(Rj)}$

D_j = j-ésimo antirango es decir $|X_{Dj}| = |X|^{(j)}$

Estadístico del test: el estadístico del test de Wilcoxon (1945), T+, es la suma de los rangos de los valores absolutos de las observaciones mayores que 0 en la muestra original. Es decir, si definimos

$$W_j = \begin{cases} 1 & \text{si } |X|^{(j)} \text{ corresponde a una observación mayor que 0} \\ 0 & \text{en caso contrario} \end{cases}$$

$$T^+ = \sum_{j=1}^{n} j W_j$$

Pero $W_j = s(X_{D_j})$, con $s(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$.

entonces, podemos expresar al estadístico en la forma

$$T^+ = \sum_{j=1}^{n} R_j \, s(X_j)$$

Observación: Si θ > 0 y la distribución simétrica se halla desplazada hacia la derecha, las observaciones positivas tienden a estar más alejadas del 0 que las negativas, entonces T+ tiende a ser grande y se rechazaría Ho.

PAQUETE 'WilcoxCV'

Este paquete proporciona funciones que actúan rápido en la selección de variables basados en el test suma de los rangos de Wilcoxon en validación cruzada o validación cruzada mediante Monte-Carlo, para usar microarray basado en clasificación binaria.

generate.cv : Genera grupos mediante validación cruzada

Genera aleatoriamente m grupos para realizar de validación cruzada m veces:

- **Uso**

generate.cv(n,m)

- **Argumentos**

n el número total de observaciones en el conjunto de datos

m el número deseado de grupos

- **Importante**

Una matriz de dimensión m x (n/m) da el número máximo de índices de las observaciones incluidas en cada grupo.

La i-esima fila da los índices de observaciones incluidas en el grupo i-ésimo. Si los m grupos no son exactamente igual tamaño, la última columna incluye uno o varios de ceros.

EJEMPLO:

```
library(WilcoxCV)

#genera 10 grupos para un conjunto de datos de tamaño 95.

generate.cv(n=95,m=10)
[1] 10
[1] 10
[1] 10
[1] 10
[1] 10
[1] 9
[1] 9
[1] 9
[1] 9
[1] 9
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	4	5	7	20	25	26	39	42	53	93
[2,]	1	27	40	55	57	61	65	68	80	92
[3,]	15	16	35	44	74	76	84	85	89	91
[4,]	2	9	10	11	33	45	47	60	82	94
[5,]	8	24	28	37	46	56	59	70	73	87
[6,]	17	49	50	58	64	66	71	75	77	0
[7,]	14	18	21	23	34	38	43	51	83	0
[8,]	3	6	19	36	67	72	78	79	86	0
[9,]	12	13	29	31	41	48	52	63	95	0
[10,]	22	30	32	54	62	69	81	88	90	0

generate.split : Genera divisiones aleatorias en el aprendizaje y en conjuntos de datos de prueba.

La función generate.split genera *niter* divisiones aleatorias en el aprendizaje y en conjuntos de datos de prueba para su uso en Monte-Carlo de validación cruzada (MCCV).

- **Uso**

generate.split (niter, n, ntest)

- **Atributos**

niter El número de iteraciones (número de partes en el aprendizaje y partes de los conjuntos).

n El número total de observaciones en el conjunto de datos.

ntest El número de observaciones en los conjuntos de prueba.

- **Detalles**

Esta función está pensada para su uso en Monte-Carlo de validación cruzada (MCCV).

- **Importante**

Una matriz de dimensión niter x ntest da los índices de las observaciones incluidas en los conjuntos de prueba. La i-ésima fila da los índices de las ntest observaciones incluidas en el conjunto de prueba para la i-ésima iteración MCCV.

EJEMPLO:

```
Library(WilcoxCV)

#Genera 50 divisiones con relación 2:1 para el conjunto de datos incluyendo 90 observaciones

generate.split(niter=50, n=90, ntest=30)
```

wilcox.selection.split : Wilcoxon-based selecciona variables mediante validación cruzada (CV) y mediante validación cruzada de Monte-Carlo (MCCV).

La función wilcox.selection.split ordena las variables mediante el test Wilcoxon de suma de rangos para todas las iteraciones CV o MCCV.

- **Uso**

wilcox.selection.split (x,y, split, algo=”new”, pvalue=FALSE)

- **Atributos**

x una matriz o un data frame de tamaño n x p da la expresión de los niveles de las p variables (genes) para las n observaciones (arrays). Variables correspondientes a columnas, observaciones correspondientes a filas.

y un vector de longitud n da la clase del número de miembros para las n observaciones(arrays). **y** puede ser numérico o un factor pero debe ser codificado como 0,1.

split una matriz *niter x nest* da los índices de las *ntest* observaciones incluidas en cada uno de de las *niter* de los conjuntos de prueba, como los generados por las funciones anteriormente explicadas. La fila i-ésima de Split da los índices de las observaciones incluidas en el conjunto de datos de prueba para la i-ésima división iterada aleatoriamente.

algo “new” o “naive”. Si type=”new”, nuevo método. Si type=”naive”, los resultados son obtenidos tras recorrer la función Wilcox.test *niter* veces.

- **Detalles**

El estadístico suma de los rangos de Wilcoxon es definido como la suma del rango de X-rangos de observaciones con y=0. El test de la suma del rango Wilcoxon es equivalente al test Mann-Whitney. Está implementado en la función wilcox.test.

En el contexto de CV o MCCV, wilcox.selection.split calcula el estadístico de la suma de los rangos Wilcoxon para cada iteración y para cada variable. En cada iteración, un sujeto de las n observaciones es excluido del conjunto de datos y este será considerado como el conjunto de datos de prueba. Los índices de la observación considerada como el conjunto de prueba para cada para cada iteración está dando el *split* en la matriz de dimensión niter x ntest.

- **Importante**

Ordering Split.

Una matriz niter x p da los índices de los genes ordenados por el p-valor. Por ejemplo, la primera columna de *ordering.split* da los índices de las variables con el pvalor más bajo en cada una de las iterativas divisiones aleatorias, la segunda columna de *ordering.split* da los índices de las variables con el segundo p-valor más bajo en cada una de las iterativas divisiones aleatorias. Para la i-esima iteración , el índice de las 50 mejores variables están en las 50 primeras columnas de la fila i.

pvalue.split

Devuelve solamente si pvalue=TRUE. Una matriz niter x p de valores. El elemento de la i-esima fila y la j-esima columna es el p-valor de la variable j en la iteración i-esima.

EJEMPLO:

```
#Generamos un conjunto de datos
x<-matrix(rnorm(1000),100,10)
y<-sample(c(0,1), 100, replace=T)

#Generamos 50 divisiones MCCV con proporción 2:1 para el conjunto de datos incluyendo 90 observaciones
div<-generate.split(niter=50, n=90, ntest=30)
#calculamos la suma del rango del estadístico Wilcoxon para las 50 interacciones
solucion<-wilcox.selection.split(x=x,y=y, split=div, algo="new", pvalue=T)
attach(solucion)
ordering.split
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	6	9	2	3	4	8	10	1	7	5
[2,]	6	3	2	10	9	4	7	5	8	1
[3,]	6	9	8	2	7	10	3	5	1	4
[4,]	1	6	2	9	7	8	4	10	3	5
[5,]	6	3	8	1	9	10	7	4	2	5
[6,]	6	3	2	9	7	4	8	5	1	10
[7,]	6	2	1	9	3	4	7	8	5	10
[8,]	2	9	10	3	8	1	6	4	7	5
[9,]	6	2	9	8	10	3	7	1	4	5
[10,]	9	3	7	6	8	10	4	1	5	2
[11,]	9	3	1	7	10	5	4	2	8	6
[12,]	9	6	8	2	1	10	5	4	3	7
[13,]	4	6	9	10	2	3	8	7	5	1
[14,]	9	2	6	3	5	8	7	4	1	10
[15,]	9	6	8	2	3	1	5	7	10	4
[16,]	6	9	7	2	8	4	10	3	5	1
[17,]	9	7	5	6	3	1	8	4	2	10
[18,]	2	10	6	7	8	3	1	5	9	4
[19,]	6	2	3	9	7	5	10	8	1	4
[20,]	2	6	9	8	7	4	1	10	5	3
[21,]	6	2	5	10	9	4	7	3	8	1
[22,]	9	6	2	8	10	4	3	7	5	1
[23,]	6	2	9	7	4	10	8	1	5	3
[24,]	6	9	2	3	8	5	10	4	1	7
[25,]	8	3	2	6	10	5	9	4	1	7
[26,]	6	9	7	2	8	3	10	4	5	1
[27,]	6	3	9	7	2	1	10	4	8	5
[28,]	6	9	2	8	1	4	3	7	5	10
[29,]	6	3	2	8	9	1	10	4	5	7
[30,]	9	10	5	6	7	8	2	1	3	4
[31,]	8	2	7	6	1	9	3	10	5	4
[32,]	6	2	9	7	5	3	4	10	1	8
[33,]	7	6	8	9	2	3	4	10	5	1
[34,]	8	2	9	3	6	10	1	4	5	7
[35,]	6	7	2	3	1	9	5	10	8	4
[36,]	6	7	4	2	9	10	5	8	3	1
[37,]	9	8	2	4	6	7	1	5	10	3
[38,]	6	9	3	8	2	7	5	1	4	10
[39,]	7	9	6	2	3	8	4	1	10	5
[40,]	6	9	2	10	4	5	3	1	8	7
[41,]	9	6	2	1	3	8	5	4	10	7
[42,]	9	2	6	8	10	7	4	1	3	5
[43,]	3	9	2	8	10	6	4	5	7	1
[44,]	2	9	8	7	6	3	10	4	5	1
[45,]	2	8	3	9	6	10	4	7	1	5
[46,]	2	6	9	1	5	3	4	7	10	8
[47,]	6	9	2	7	1	3	8	10	5	4
[48,]	6	3	9	1	2	5	8	7	10	4
[49,]	3	9	7	6	8	4	10	2	5	1
[50,]	7	2	8	3	9	1	10	5	6	4

pvalue.split


```
> pvalue.split
             [,1]        [,2]       [,3]       [,4]      [,5]        [,6]
 [1,] 0.72505101 0.226224164 0.38737500 0.40731956 0.8533858 0.065469331
 [2,] 0.97622140 0.150811270 0.11145990 0.25486706 0.5793049 0.067232416
 [3,] 0.67987639 0.294056746 0.61219155 0.87819069 0.6121915 0.029179433
 [4,] 0.05037106 0.157680558 0.86387652 0.77206317 0.9575678 0.154232630
 [5,] 0.15422971 0.663435809 0.04207261 0.47071478 0.8627477 0.040882099
 [6,] 0.66058240 0.122875388 0.04617466 0.60154227 0.6265557 0.009676390
 [7,] 0.20413580 0.122875388 0.36081045 0.39944381 0.7307209 0.036731520
 [8,] 0.48782880 0.004664122 0.31370931 0.73094649 0.9855605 0.503090323
 [9,] 0.68086640 0.068547919 0.51050613 0.85086550 0.9251035 0.049716103
[10,] 0.80318740 0.971598558 0.10147204 0.65200610 0.9243605 0.264604202
[11,] 0.42425534 0.680866405 0.29024694 0.62163295 0.6051288 0.887861511
[12,] 0.63499179 0.545007154 0.84010992 0.82159926 0.7576538 0.226080460
[13,] 0.98591968 0.246503602 0.38074954 0.07859198 0.9484055 0.124692033
[14,] 0.81419493 0.097544667 0.39751702 0.66372001 0.4807770 0.172849997
[15,] 0.46627383 0.295639916 0.34718106 0.72444331 0.4662738 0.200247430
[16,] 0.97187827 0.115344372 0.80508725 0.41747400 0.8508655 0.006638483
[17,] 0.45500465 0.791225460 0.40684709 0.66760738 0.2188935 0.343583757
[18,] 0.46213715 0.124692033 0.42028657 0.81853964 0.7197132 0.276466318
[19,] 0.84939920 0.120021354 0.13794770 0.98106348 0.4265299 0.031707814
[20,] 0.69532575 0.075047737 0.90553034 0.46910453 0.8587094 0.099019351
[21,] 0.87819069 0.045047393 0.75024527 0.40256272 0.1751659 0.016170077
[22,] 0.87819069 0.443488170 0.73242997 0.69723705 0.8227572 0.370252495
[23,] 0.91577436 0.092883951 0.97187827 0.48811975 0.9251035 0.084099306
[24,] 0.94379145 0.226147837 0.28492057 0.86008706 0.7421369 0.020617025
[25,] 0.75024527 0.178948143 0.13741817 0.70597789 0.3278138 0.262713494
[26,] 0.92034020 0.159737390 0.23240771 0.83687446 0.8460759 0.046127891
[27,] 0.50411368 0.347231357 0.16118657 0.71838457 0.8360783 0.079122767
[28,] 0.36398919 0.294056746 0.53984131 0.45764077 0.8411512 0.046324757
[29,] 0.38149432 0.053592327 0.03698179 0.84788878 0.8572903 0.030015693
[30,] 0.48811975 0.459114320 0.51808579 0.65521170 0.2086287 0.212913384
[31,] 0.43398332 0.349611734 0.69347908 0.86454059 0.7552070 0.387175599
[32,] 0.70956366 0.032830277 0.41122632 0.58245958 0.3352364 0.013693202
[33,] 0.93444296 0.274467254 0.54116896 0.75996776 0.8508655 0.079966148
[34,] 0.75103569 0.239959134 0.58074845 0.81419493 0.9344430 0.588824053
[35,] 0.36809539 0.119018998 0.30326113 0.94840550 0.7109299 0.038940692
[36,] 0.95777659 0.361867675 0.94840550 0.29776858 0.6421248 0.031779729
[37,] 0.56499467 0.175520018 0.75527162 0.23059386 0.5813170 0.264887973
[38,] 0.69238976 0.548896006 0.19541099 0.80123090 0.6747845 0.028245488
[39,] 0.78695219 0.397517021 0.48811975 0.62163295 0.9437915 0.138709978
```

wilcox.split : El estadístico de la suma del rango Wilcoxon en validación cruzada (CV) y validación cruzada de Monte-Carlo (MCCV).

La función wilcox.split calcula el estadístico de la suma de los rangos de Wilcoxon para todas las iteraciones CV o MCCV definidas por la matriz *split*

- **Uso**

wilcox.split (x,y, split, algo=”new”)

- **Atributos**

x una matriz o un data frame de tamaño n x p da la expresión de los niveles de las p variables (genes) para las n observaciones (arrays). Variables correspondientes a columnas, observaciones correspondientes a filas.

y un vector de longitud n da la clase del número de miembros para las n observaciones(arrays). **y** puede ser numérico o un factor pero debe ser codificado como 0,1.

split una matriz niter x nest da los índices de las *ntest* observaciones incluidas en cada uno de de las niter de los conjuntos de prueba, como los generados por las funciones anteriormente explicadas. La fila i-ésima de Split da los índices de las obsevaciones incluidas en el conjunto de datos de prueba para la i-ésima división iterada aleatoriamente.

algo “new” o “naive”. Si type=”new”, nuevo método. Si type=”naive”, los resultados son obtenidos tras recorrer la función Wilcox.test *niter* veces.

- **Detalles**

El estadístico de la suma de los rangos de Wilcoxon es definido como la suma del rango de X-rangos de observaciones con y=0. El test de la suma del rango Wilcoxon es equivalente al test Mann-Whitney. Está implementado en la función wilcox.test.

En el contexto de CV o MCCV, wilcox.selection.split calcula el estadístico de la suma de los rango Wilcoxon para cada iteración y para cada variable. En cada iteración, un sujeto de las n observaciones es excluido del conjunto de datos y este será considerado como el conjunto de datos de prueba. Los índices de la observación considerada como el conjunto de prueba para cada para cada iteración está dando el *split* en la matriz de dimensión niter x ntest.

- **Importante**

Wilcox. Split.

Un numérico vector de longitud niter el cual su i-esima componente da el estadístico de la suma del rangp Wilcoxon obtenido en la i-ésima interacción.

EJEMPLO:

```
#Generamos un conjunto de datos
x<-rnorm(100)
y<-sample(c(0,1), 100, replace=T)

#Generamos 50 divisiones MCCV con proporción 2:1 para el conjunto de datos incluyendo 90 observaciones
div<-generate.split(niter=50, n=90, ntest=30)

#calculamos la suma del rango del estadístico Wilcoxon para las 50 interacciones
wilcox.split(x=x,y=y, split=div, algo="new")
```

```
[1] 1170 1232 1149 1146 1124 1188 1203 1103 1036 1146 1142 1227 1091 1134 1100
[16] 1167 1205 1229 1339 1167  984 1169 1188 1089 1013 1265 1141 1191 1053 1334
[31] 1197 1355 1239 1214 1249 1035 1082 1073 1080 1050 1159 1121 1309 1113 1183
[46] 1152 1116 1085  952 1223
```