Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Inteligencia artificial
Recibido: 18/10/2012 | Aceptado: 25/10/2013 | Publicado: 10/12/2013

Una revisión a algoritmos de selección de atributos que tratan la redundancia en datos microarreglos

A review of feature selection algorithms that treat the microarray data redundancy

Roxana Pérez Rubido

Departamento Ciencias Básicas. Facultad 7. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370

rubido@uci.cu

Resumen

En los últimos tiempos, el análisis de la redundancia en los algoritmos de selección de atributos en el aprendizaje automático, se ha convertido en una constante. Estudios han demostrado que los porcientos de predicción al eliminar estos atributos son mejores que los obtenidos en los casos donde no se hace. Además, al descartarlos se disminuye la complejidad temporal del clasificador al tener menos datos que procesar. En la actualidad, los algoritmos han evolucionado en ese sentido y tratan la redundancia de diferentes formas y con diferentes criterios. El principal objetivo del presente trabajo es presentar diferentes criterios de evaluación para tratar la redundancia en datos microarreglos de ADN. En el estudio se aplicaron los métodos análisis y síntesis, histórico-lógico e inductivo-deductivo. Se realizó una revisión bibliográfica de artículos publicados desde la década del 90 que presentan algoritmos para seleccionar atributos y que tienen en cuenta la dependencia entre ellos. En el artículo se describen de forma general los pasos, el criterio empleado en el análisis de la redundancia y algunas de las ventajas y desventajas de los mismos.

Palabras clave: Algoritmos filtros, análisis de redundancia, criterios de evaluación, selección de atributos.

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

Abstract

In recent times, the redundancy analysis in attribute selection algorithms in machine learning has become a constant. Studies have shown that the percentages of prediction, after removing these attributes, are better than the cases where it is not. Furthermore, by excluding it from data set, the temporal complexity of the classifier is reduced because it has less data to process. In the actually, the algorithms have evolved in this regard and treat redundancy in different ways and with different criteria. The main aim of this review is to present the different evaluation criteria to address data redundancy in ADN microarrays. The study applied analysis-synthesis, historic-logical and inductive-deductive methods. We conducted a literature review of articles published since the 90's which contain algorithms to select attributes and take into account the dependency between them. The article describe a general way, his steps, the criterion used in the analysis of redundancy and some of its advantages and disadvantages.

Keywords: Analysis of redundancy, evaluation criteria, feature selection, filter algorithms.

Introducción

El análisis de los datos microarreglos de ADN genera un gran cúmulo de información, por lo que constituye un reto para el aprendizaje automático. Su gran dimensionalidad, dada por contener un número significativo de atributos irrelevantes para la clasificación o redundantes en el contexto de otros, afecta el aprendizaje en términos de precisión y de complejidad computacional. Resulta entonces una necesidad reducir el conjunto de datos, seleccionando un grupo de atributos a través de la eliminación de aquellos que no son útiles para la tarea de predicción. Los métodos (algoritmos) de selección de atributos pueden, de forma general, ser divididos en dos categorías: envolvente (wrapper) y filtro (Bonev, 2010; Guyon and Elisseeff, 2003; Kohavi and John, 1997). Los primeros son dependientes del clasificador, pues usan la precisión estimada de un algoritmo de clasificación para medir la bondad de un subconjunto de atributos en particular. Además, son computacionalmente costosos para conjuntos de datos con una gran dimensionalidad, pues realizan la búsqueda a través del espacio de subconjuntos de atributos. Los segundos, a diferencia de los primeros, son independientes del clasificador y se basan en las características generales de los datos de entrenamiento para la selección. Las funciones de evaluación utilizadas están basadas en diversos criterios tales como el coeficiente de correlación de Pearson, la información mutua, la incertidumbre simétrica o en heurísticas que combinan varios de estos criterios u otros. Por ser rápidos y con un costo computacional menor al de los wrapper, son a menudo los adoptados para reducir la dimensión en datos microarreglos.

En la literatura se encuentran disímiles algoritmos filtro (Saeys *et al.*, 2007), los primeros que surgieron (univariados) evalúan a cada gen (atributo) de forma individual, sin importar la relación de dependencia existente entre ellos.

Ordenan a los genes a partir de su poder discriminativo con respecto a la clase o el grado de relevancia individual y finalmente seleccionan los primeros k elementos. Pese a ser eficientes, computacionalmente hablando, pues su complejidad de tiempo es lineal y está en términos de la dimensión del conjunto de datos, presentan dos desventajas principales: 1) se necesita que se posea conocimientos del dominio en el que se esté trabajando para determinar el número de elementos a seleccionar (umbral) y 2) no se eliminan los atributos redundantes.

Ha sido demostrado en varios estudios (Blum and Langley, 1997; Li and Yang 2002; Xiong *et al.*, 2001; Yang and Pedersen, 1997) que la combinación de genes relevantes no siempre hace del subconjunto seleccionado el mejor, pues puede que estos compartan la misma información, lo que los hace redundantes (correlacionados). Los conjuntos que contienen datos correlacionados entre sí son menos abarcadores; o sea, no representan todas las características del conjunto original, lo que afecta la precisión de la predicción. Por ello, varios algoritmos en la actualidad (multivariados) (Battiti, 1994; Biesiada and Duch, 2008; Brown, 2009; Ding and Peng, 2005; Fleuret, 2004; Hall, 1999; Meyers *et al*,. 2005; Yu and Lei, 2004; Zheng and Kwoh, 2011), a diferencia de los primeros, tienen en cuenta la correlación entre atributos y eliminan a aquellos que no aportan nueva información al proceso de clasificación. Varios de ellos, siguen dependiendo de la selección de un valor umbral para decidir cuántos atributos seleccionar y además, tienen mayor complejidad de tiempo computacional en comparación con los primeros, aunque siguen siendo más rápidos que los wrapper en datos de grandes dimensiones. En este artículo se presenta una selección de los algoritmos filtros que tratan la redundancia, publicados en las dos últimas décadas, divididos según la dirección de la búsqueda que emplean, haciendo énfasis en las funciones de evaluación empleadas para el análisis de la redundancia y algunas de sus ventajas y desventajas.

Materiales y métodos

Los métodos de investigación empleados son el analítico-sintético, histórico-lógico e inductivo-deductivo. El método analítico-sintético se empleó para examinar los elementos de los algoritmos de selección de atributos y definir los esenciales para la investigación. El método histórico-lógico se utilizó para determinar las distintas etapas de los algoritmos descritos y la evolución de las funciones de evaluación siguiendo su lógica interna y el método inductivo-deductivo para determinar, a partir de las definiciones y conceptos existentes, los que se ajustan a la investigación y replantearlos o para, a partir de elementos singulares llegar a proposiciones generales.

La selección de los algoritmos se realizó de forma tal que fuera una muestra heterogénea a partir de las funciones de evaluación y la dirección de la búsqueda que emplean. Además, se tuvo en cuenta que los resultados de la predicción fueran semejantes (no exactamente iguales) para una complejidad temporal similar.

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

Resultados y discusión

Notación básica

Para un mejor entendimiento del contenido del artículo se declara la notación empleada en la descripción de los algoritmos. La letra C es la variable de salida (clase objetivo), el conjunto de las variables de entrada (conjunto de datos original) se denota por F y el conjunto de los atributos seleccionados al aplicar algún criterio por G. En la **Tabla 1. Notación básica** se muestra un resumen de la notación empleada.

Tabla 1. Notación básica.

Notación	Descripción
$F = \{f_i : i = 1..n\}$:	El conjunto original de n atributos.
$S_l = \{s_i : i \leq n\}, S_l \subseteq F$:	Un subconjunto de F.
$\|F\| = n$:	Cantidad de atributos en el conjunto F.
$\|S_l\| = k$:	Cantidad de atributos en el conjunto S_l.
N:	Número de muestras (experimentos).
$G = \{g_i : i = 1..m; m \leq n\}, G \subseteq F$:	El conjunto de los m atributos ya seleccionados.
$Q_i = \{q_j : j \leq n\}, Q_i \subseteq F \backslash f_i$:	La manta de Markov para el atributo f_i.
$\|G\| = m$:	Cantidad de atributos en el conjunto G.
$\|M_i\| = r$:	Cantidad de atributos en el conjunto M_i.

Algoritmos filtros. Atributos relevantes y redundantes

Los algoritmos filtro se basan en las características de los atributos para seleccionar el subconjunto que contenga aquellos que son relevantes (la relevancia se establece al aplicar diferentes criterios) para la tarea de clasificación. Al igual que cualquier algoritmo de selección de atributos, parten del conjunto de datos completo F, o del conjunto vacío ϕ o bien de un conjunto $S \subset F$ cualquiera. Iterativamente analizan otros subconjuntos, una vez terminado con el primero, y se detendrán cuando ya no queden subconjuntos por analizar o se cumpla la condición de parada

establecida en el algoritmo. La búsqueda del próximo subconjunto depende de la dirección de la misma: hacia adelante (forward), hacia atrás (backward), aleatoria (random), entre otras.

Para determinar si un atributo es relevante para la tarea de clasificación o si es redundante con respecto a otros existen diferentes criterios. Los mismos son de disímiles naturalezas: estadísticos como el coeficiente de correlación de Pearson (Meyers *et al.*, 2005), pruebas Kolmogorov-Smirnov (Biesiada and Duch, 2008), cálculos relacionados con la entropía (Frey and Fisher, 2003; Khinchin, 1957; Shannon and Weaver, 1963) como la información mutua, información mutua condicional, entre otros.

La definición más comúnmente usada de atributo relevante es la planteada en (John *et al.*, 1994), donde se especifican dos tipos: fuerte y débil. Un atributo es fuertemente relevante si al eliminarlo del conjunto de datos afecta la precisión del clasificador, pues aporta información que ningún otro tiene, por lo que son atributos necesarios en el subconjunto óptimo. Un atributo es débilmente relevante si no es fuertemente relevante pero bajo ciertas condiciones aporta información nueva, no siempre es necesario pues su información puede ser suministrada por un conjunto de atributos.

La redundancia normalmente es definida en términos del grado de dependencia (correlación) que existe entre los atributos, por lo que se dice que dos atributos son redundantes entre sí, si están altamente correlacionados. Se distinguen dos tipos de correlación: lineal y no lineal. En la literatura es común encontrar este análisis entre pares de atributos(Brown, 2009; Hall, 1999) sin tener en cuenta la dependencia que puede existir entre grupos de atributos (complementariedad entre atributos (Meyer *et al.*, 2008)).

Por lo general, los algoritmos no tratan la redundancia de forma independiente, sino que a partir de una misma función de evaluación tratan de seleccionar los atributos más relevantes para la clase, pero menos redundantes con respecto a otros atributos.

A continuación se describen algunos de los algoritmos filtros que tratan la redundancia a partir de diferentes criterios y tipos de datos, presentados en orden cronológico.

Algoritmos filtros con búsqueda forward

Los algoritmos que utilizan una búsqueda forward en el espacio de subconjuntos comienzan con el conjunto vacío y van adicionando los "mejores" atributos escogidos del conjunto F de acuerdo al criterio de evaluación en cada iteración. La búsqueda puede detenerse por diversas razones: se seleccionaron la cantidad de atributos predefinida, no se alcanzan resultados mejores pasadas algunas iteraciones, no quedan subconjuntos por analizar, entre otras.

El **algoritmo MIFS** (BATTITI 1994) (del inglés, *Mutual Information Features Selection*) es el primero, de varios, que intenta balancear la relevancia con la redundancia. Selecciona a los atributos relevantes para la predicción con la

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

menor correlación con otros atributos. El algoritmo calcula el valor $I(C,f_i)$ para cada atributo $f_i \in F\backslash G$ y selecciona al de mayor información mutua como el primer elemento del subconjunto de atributos seleccionados G. Luego para el próximo f_i se calcula la información mutua de (f_i,g_j) y se seleccionan los m atributos f_i que maximizan el criterio MIFS. Este criterio incluye el término $I(C,f_i)$ para garantizar la relevancia del atributo pero introduce una penalidad $\beta \sum_{g_j} I(f_i,g_j)]$, para forzar que exista baja correlación.

Definición (criterio MIFS): Sea$f_i \in F\backslash G$un atributo, la condición MIFS es:

$$MIFS = \arg max_{f_i} [I(C,f_i) - \beta \sum_{g_j} I(f_i,g_j)],$$

donde β es un parámetro de peso que es configurable, y es quien regula la importancia relativa de la información mutua entre el atributo candidato y los ya seleccionados, con respecto a la información mutua con la clase. Si toma valor 0 la expresión resultará en el cálculo de la relevancia individual, si toma un valor grande denotará mayor énfasis en la reducción de la correlación entre los atributos.

Ventajas

Se logra un balance entre la importancia (relevancia) individual ($I(C,f_i)$) y la correlación con el resto de los atributos ya seleccionados ($I(f_i,g_j)$). No se hacen suposiciones sobre la distribución de los datos.

Desventajas

Sufre la dificultad de calcular apropiadamente el valor del parámetro β. Depende de un término como umbral para decidir la cantidad de elementos a seleccionar. No tiene en cuenta la interacción entre grupos de atributos.

El **algoritmo EWUSC** (del inglés, *Error-Weighted, Uncorrelated Shrunken Centroid*) (Yeung and Bumgarner, 2003), es una modificación del algoritmo USC (del inglés, *Uncorrelated Shrunken Centroid*), el cual constituye un método integrado de selección y clasificación de atributos. Se basa en las estimaciones del error o variabilidad de las mediciones repetidas. El algoritmo, para cada umbral de contracción Δ, selecciona los atributos relevantes formando un conjunto S_Δ. Los atributos relevantes serán aquellos que posean al menos una diferencia relativa ponderada $d_{s_ic} > 0$ (diferencia entre la clase centroide y todos los centroides, estandarizado por la desviación estándar dentro de la clase del atributo i) sobre todas las clases. Se ordenan en forma descendiente, a partir del valor de la d_{s_ic}. Luego, se calcula la correlación pairwise (por pareja) ponderada basada en el error entre cada par de atributos $(s_i,s_j) \in S_\Delta$ y se eliminan aquellos atributos $s_j$cuyo valor de correlación $\tilde{\rho}_{s_i,s_j}$sea mayor que un valor ρ_0. Los valores de Δ y $\rho_0$están

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

determinados por la validación cruzada de manera tal que el número de errores de clasificación se reduce al mínimo en el conjunto de entrenamiento.

Definición (correlación basada en el error ponderado): Sea $\sigma_{s_i e}$ el error estimado del nivel de expresión del atributo s_i bajo el experimento e. La correlación basada en el error entre el par de atributos s_i y s_j se define como:

$$\tilde{\rho}_{s_i,s_j} = \frac{\sum_{e=1}^{n} \frac{(D(s_i,e) - \tilde{\mu}_i)}{\sigma_{s_i e}} \frac{(D(s_j,e) - \tilde{\mu}_j)}{\sigma_{s_j e}}}{\sqrt{(\sum_{e=1}^{n} \frac{D(s_i,e) - \tilde{\mu}_i}{\sigma_{s_i e}})^2 (\sum_{e=1}^{n} \frac{D(s_j,e) - \tilde{\mu}_j}{\sigma_{s_j e}})^2}}$$

donde $\tilde{\mu}_{s_i} = \sum_{e=1}^{n} \frac{D(s_i,e)}{\sigma_{s_i e}} / \sum_{e=1}^{n} \frac{1}{\sigma_{s_i e}}$ es el nivel de expresión promedio ponderado del atributo i y $D(s_i,e)$ representa el nivel de expresión promedio sobre las mediciones repetidas para un atributo s_i bajo el experimento e.

Ventajas

El algoritmo analiza la redundancia luego de la relevancia, reduciendo la dimensión del conjunto de datos relevantes al eliminar aquellos que no aportan información nueva sobre la clase. Puede ser usado en problemas con múltiples clases y no hace suposiciones sobre la distribución de los datos. Combina, para determinar el análisis de la redundancia una búsqueda forward con una eliminación backward. Trabaja con datos continuos.

Desventajas

Depende de las estimaciones de error o variabilidad, pues hace que mejore la estabilidad de los rasgos con estimaciones de error. No se obtienen buenos resultados cuando las clases no están bien separadas; o sea, cuando el ruido biológico no es pequeño o cuando la razón ruido-señal es baja.

El **algoritmo CMIM** (del inglés, *Conditional Mutual Information Maximization*) (Fleuret, 2004) se basa en la información mutua condicional y no trata la redundancia de forma independiente a la relevancia. El algoritmo selecciona un nuevo atributo $f_i \in F \backslash G$ solamente si aporta información nueva sobre la clase, información que no esté presente en G. El primer atributo seleccionado es aquel que tiene una información mutua máxima con respecto a la clase. Iterativamente se seleccionan los atributos f_i que maximizan su información mutua con la clase a predecir, condicionada a cualquier atributo g_j ya seleccionado.

Definición (CMIM): Sea $f_i \in F \backslash G$ un atributo, se define el criterio CMIM como:

$$CMIM = \arg max_{f_i} \{ min_{g_j} [I(f_i, C | g_j)] \}$$

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

Ventajas

El algoritmo logra un balance entre el poder individual de cada atributo y su independencia al compararlos con los atributos ya seleccionados. Además, calcula densidades bivariadas y tiene en cuenta la interacción entre grupos de atributos. Trabaja con datos binarios. Realiza el análisis de la redundancia junto con el paso de selección lo que provoca que este último aumente su complejidad al realizar la búsqueda en un espacio de gran dimensionalidad.

Desventajas

Depende de un término como umbral para determinar cantidad de atributos a seleccionar. No garantiza la selección de todos los atributos que interactúan con los pertenecientes a G. Puede ocurrir que un atributo tenga una alta información mutua condicional con respecto a otros atributos ya seleccionados (complementariedad) pero no necesariamente será la mayor información mutua condicional.

El **algoritmo FCBF** (del inglés, *Fast Correlation Based Filter*) (Yu and Lei, 2004) combina un método ranking con el análisis de la redundancia. El análisis de los atributos redundantes se hace de forma independiente al de la relevancia, determinándose a partir de los atributos seleccionados como relevantes. La redundancia se define a partir de las mantas de Markov:

Definición (atributo redundante-mantas de Markov): Un atributo $f_i \in F$ es **redundante** si y solo si (ssi) es débilmente relevante; o sea, aquel que es independiente condicionalmente del resto de los atributos $(F \backslash f_i)$ pero no de un subconjunto de ellos ($P(f_i, C | F \backslash f_i) = P(C | F \backslash f_i)\ pero\ \exists\ S_i \subset F \backslash f_i, P(f_i, C | S_i) \neq P(C | S_i)$); y tiene una manta de Markov aproximada M_i dentro de F.

El algoritmo, en un primer paso, selecciona los atributos (la cantidad de atributos está condicionada por un valor umbral) más relevantes con respecto a la clase a partir del valor $I(f_i, C)$. En un segundo paso se analiza si cada atributo $s_i \in S_l$ (S_l conjunto de los atributos más relevantes) es redundante con respecto a los seleccionados en G y en caso positivo se elimina. Este paso se repite hasta que no queden atributos redundantes.

El análisis de la redundancia se divide en 3 pasos: primero, determinar un atributo predominante, es decir, un atributo que no tenga ninguna manta de Markov aproximada en el conjunto S_i; segundo, eliminar todos los atributos para los cuales este forme una manta de Markov aproximada y tercero repetir los pasos anteriores hasta que no haya más atributos predominantes.

Definición (Manta de Markov aproximada): Para dos atributos relevantes $f_i \in F, f_j \in F$, f_j forma una manta de Markov aproximada para f_i ssi el valor de relevancia de f_j es mayor o igual que el valor de relevancia de f_i con

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

respecto a la clase y además, el grado de correlación entre ellos es mayor que el valor de relevancia de f_i con respecto a la clase.

Ventajas

El criterio para medir el nivel de relevancia y de correlación es la incertidumbre simétrica lo que evita la parcialidad de la información mutua hacia los atributos multi-evaluados, penalizando los atributos con grandes entropías. Trabaja con datos discretos. Realiza la selección y el tratamiento de la redundancia no simultáneamente. Combina, para determinar el análisis de la redundancia una búsqueda forward con una eliminación backward.

Desventajas

La selección del conjunto de los atributos relevantes se basa en un umbral predefinido δ buscando los atributos que están más correlacionados con la clase. Es un algoritmo rápido pero puede eliminar atributos redundantes que están fuertemente correlacionados con la clase, en situaciones donde la dependencia entre atributos ocurre solamente condicionalmente sobre esta, debido a que está fundamentado en dos funciones de costo.

El **algoritmo CMIFS** (del inglés, *Conditional Mutual Information-based Feature Selection*) (Cheng *et al.*, 2011) se basa en la información mutua condicional y en el criterio *atributo redundante para la clasificación* (FCR), el cual define la información de redundancia de un atributo que está relacionada con la clasificación. Se seleccionan los atributos $f_i \in F \backslash G$ que no sean FCR y que maximicen la siguiente relación de recurrencia:

Definición (criterio CMIFS) Sea el atributo g_1 el primero, del conjunto G y $f_i \in F \backslash G$, se define el próximo atributo a seleccionar como:

$$CMIFS_{n+1} = \arg max_{fi} \{I(C, f_i|g_1) - [I(g_n, f_i|g_1) - I(g_n, f_i|C)]\}$$

Para detectar los atributos FCR se auxilian de un parámetro δ. El valor de δ debe estar en un rango de valores razonable, para evitar que sean eliminados atributos como FCR irracionalmente. El primer atributo que es seleccionado es el más relevante para la clase. Luego se va construyendo el conjunto G paso a paso, eliminándose primero de F todos los atributos que cumplen con el criterio FCR con respecto a los atributos en G y luego adicionando el atributo f_i que maximiza el criterio CMIFS (usan una fórmula de estimación aproximada).

Definición (criterio FCR): Un atributo $f_i \in F \backslash G$ es un FCR de G si $I(f_i, G) > 0$ y $I(C, f_i|G) = 0$.

Ventajas

Se tiene en cuenta la interacción entre grupos de atributos. Se combina el cálculo de la redundancia con la tarea de clasificación, lo que puede disminuir la probabilidad de confundir a un atributo importante con un atributo redundante

Revista Cubana de Ciencias Informáticas
Vol. 7, No. 4, octubre-diciembre, 2013
ISSN: 2227-1899 | RNPS: 2301
Pág. 16-30
http://rcci.uci.cu

durante el proceso de búsqueda. La existencia de atributos que cumplen el criterio FCR tiene poco impacto en los resultados del criterio de selección de subconjunto de atributos, sin embargo ayudan a reducir el tiempo, pues permite que sean eliminados atributos antes de hacer una nueva búsqueda a través del espacio de subconjuntos candidatos.

Desventaja

Depende de un valor umbral.

Algoritmos con búsqueda backward

Los algoritmos que utilizan una búsqueda backward en el espacio de subconjuntos comienzan con el conjunto completo F y van eliminando los "peores" atributos escogidos de acuerdo al criterio de evaluación en cada iteración. La búsqueda puede detenerse por diversas razones: el conjunto de atributos posee la cantidad de atributos predefinida, no se alcanzan resultados mejores pasadas algunas iteraciones, entre otras.

El **algoritmo KS** (Koller and Sahami, 1996) utiliza la idea de independencia condicional para reducir el conjunto de datos, auxiliándose del cálculo de la entropía cruzada y utilizando una eliminación hacia atrás (backward). El algoritmo comienza calculando el coeficiente de correlación de Pearson entre cada par de atributos (f_i, f_j), $f_i \in F, f_j \in F, i \neq j$, para, a partir de este, conformar para cada f_i su manta de Markov M_i con los r atributos f_j más correlacionados con él (los r atributos con mayor coeficiente de correlación). Luego se determina el atributo f_ia eliminar a partir de la heurística empleada para estimar mantas de Markov aproximadas y se detiene cuando hayan sido eliminados un número especificado de atributos. La heurística se basa en el cálculo de la entropía cruzada:

$$\delta_G(f_i|M_i) = \sum_{q_j, f_i} P(M_i = q_j, F = f_i) * D\left(P(C|M_i = q_j, F = f_i), P(C|M_i = q_j)\right),$$

siendo $D(\mu, \sigma) = \sum_{x \in \Omega} \mu(x) log \frac{\mu(x)}{\sigma(x)}$, donde Ω es el espacio de probabilidades, μ la distribución real, σ distribución aproximada. $D(\mu, \sigma)$ mide la magnitud del error que se comete al usar σ como sustituto de μ.

La heurística consiste en tomar el conjunto M_i con el menor valor $\delta_G(f_i|M_i)$, siendo este f_i eliminado ($M_i$una manta de Markov aproximada). El fundamento está en que si M_i es una manta de Markov para f_i el valor de $\delta_G(f_i|M_i) = 0$.

Ventajas

El algoritmo permite descartar aquellos atributos que no se acerquen a la distribución del conjunto F, por ser irrelevantes o redundantes y trabaja con datos discretos. Además, basándose en que el cálculo de las probabilidades

condicionales puede ser computacionalmente costoso, se plantean criterios que producen resultados aproximados (la heurística) a un menor costo.

Desventajas

Depende de dos parámetros: cantidad de atributos que permanecerán en el conjunto (o cantidad de atributos a eliminar) y el tamaño de la manta de Markov. Puede sufrir el problema de buscar a través de un subconjunto requerido en el paso de generación de subconjuntos, lo que puede provocar una complejidad temporal $O(n^2)$.

El **algoritmo CFS** (del inglés, *Correlation Features Selection*) (Hall 1999) clasifica subconjuntos de atributos de acuerdo a los valores de una función de evaluación heurística CFS_{S_l} basada en la correlación (correlación atributo-clase, atributo-atributo). No trata a la redundancia de forma independiente; sino que en una misma función se determinan los atributos altamente relevantes para la clasificación con respecto a la clase pero con poca o ninguna correlación con otros atributos ya seleccionados. Además, se asume que los atributos son condicionalmente independientes de otros. El algoritmo calcula las correlaciones atributo-atributo y clase-atributo y luego busca en el espacio de subconjunto de atributos, reportando el subconjunto l con mejor puntuación CFS_{S_l}.

Definición (selección de atributos basada en la correlación): Sean $\bar{r}_{cs_i}$ la correlación promedio entre el atributo s_i y la clase $c \in C$ y $\bar{r}_{s_i s_j}$ la correlación promedio entre un par de atributos, el criterio de selección se define como:

$$CFS_{S_l} = \frac{k\bar{r}_{cf_i}}{\sqrt{k + k(k-1)\bar{r}_{f_i f_j}}}$$

El numerador puede ser visto como un indicador de cuán predictivo de la clase es un conjunto de atributos y el denominador como un indicador de cuánta redundancia existe entre los atributos.

Ventajas

No necesita se predefina un término como umbral ni la cantidad de elementos a seleccionar. Pueden usarse diferentes criterios, según el problema de clasificación o la experiencia del investigador, para calcular las correlaciones promedio entre atributo-clase y atributo-clase.

Desventajas

No tiene en cuenta la cooperación entre atributos. Cuando existe dependencia condicional y esta es fuerte, el algoritmo puede fallar provocando que no sean seleccionados todos los atributos relevantes necesarios.

Entre los algoritmos descritos se enuncian desventajas que pueden afectar la ejecución del clasificador en términos de precisión. Aquellos que dependen de un umbral (MIFS, CMIM, CMIFS) ya sea para determinar el tamaño de los subconjuntos (o de las mantas de Markov) como para detener la ejecución o determinar el valor de la penalización en los atributos redundantes, pese a que su objetivo en la mayoría de los casos es evitar que se haga una búsqueda exhaustiva en el conjunto F, sufren la dificultad de seleccionar (calcular) adecuadamente su valor, y para esto se necesita dominar el problema y la naturaleza de los datos. Por ejemplo, si el valor es pequeño, en el caso del umbral para determinar el tamaño del subconjunto óptimo, se pueden obtener subconjuntos grandes, aumentando el costo computacional, mientras que si es muy grande se obtienen conjuntos pequeños, disminuyendo bastante el poder predictivo del clasificador, por lo que, si no se logra un valor adecuado los resultados obtenidos al final de la clasificación no serán satisfactorios ni fiables. Varios de estos algoritmos asumen el cálculo bivariado para evitar los costosos cálculos multivariados (muy costosos en datos de grandes dimensiones como los datos microarreglos de ADN); debido a esto no analizan la complementariedad entre grupos de atributos eliminando en muchos casos atributos que analizados con respecto a otro (bivariado) es redundante o irrelevante pero que cuando están junto a otros aportan información muy valiosa para la tarea en cuestión.

La mayoría de ellos evalúan la redundancia independientemente del problema de clasificación en cuestión, o sea, analizan si dos atributos son redundantes entre sí y de serlo eliminan uno sin importar si la información que aporta es relevante o no para la tarea de clasificación, lo que supone pérdida de información importante y puede traer como consecuencia la degradación de la predicción. Esto ocurre cuando la información redundante entre dos importantes atributos es (raramente) relativa a la clase objetivo correspondiente, por lo que ninguno puede ser ignorado. Además, como el atributo candidato es comparado con cada uno de los ya seleccionados, uno por uno, se introducen algunos cálculos redundantes. La mayoría usa una sola dirección de búsqueda y otros combinan la búsqueda forward con una eliminación backward, ambos tipos tienen resultados adecuados dados por la condición de parada que implementan; pero la combinación propicia que no solo se evite el análisis completo de la correlación entre atributos sino también que alcance mayor eficiencia que si se hiciera una búsqueda pura forward o una eliminación backward.

A partir del análisis de los algoritmos se puede decir que un criterio adecuado para el análisis de la redundancia es la información mutua pues permite calcular cuanta información comparten, no solo un par de atributos sino también un conjunto de atributos, además de no hacer suposiciones sobre la naturaleza de los datos.

Conclusiones

En el presente artículo se revisaron los algoritmos filtros MIFS, EWUSC, CMIM, FCBF, CMIFS, KS Y CFS. Se describieron de forma general los pasos que siguen para la selección; además, se señalaron algunas de sus ventajas y desventajas. La vía que emplean para seleccionar los atributos no coincide en todos los algoritmos; algunos a través de una sola función eligen al que mejor balance presenta entre la redundancia en el contexto de otros y el poder discriminativo. Otros, seleccionan un conjunto de atributos relevantes y luego realizan el análisis de la redundancia. Entre los criterios usados en los algoritmos, el más común es el cálculo de la información mutua, la cual permite capturar las dependencias que aportan información sobre la clase, sin tener en cuenta aquellas dependencias entre atributos que son irrelevantes para la clasificación. Aunque está parcializada hacia atributos multievaluados, en la literatura se encuentran varias formas de normalizar su valor, por ejemplo penalizando aquellos que mayor valor de entropía presenten. Los más recientes hacen uso de la información mutua condicional, con el fin de no solo analizar la redundancia entre un par de atributos, sino entre grupos de atributos, lo que propicia mayor eficiencia en la detección de redundancias y mayor reducción del conjunto de datos original.

Referencias

- BATTITI, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 1994, 5(4): 537-550.
- BIESIADA, J. and W. DUCH A Kolmogorov-Smirnov Correlation-Based Filter for Microarray Data. *Neural Information Processing*, 2008: 285-294.
- BLUM, A. and P. LANGLEY Selection of relevant features and examples in machine learning. *Artificial Intelligence.* 1997, 97(1-2): 245-271.
- BONEV, B. I. *FEATURE SELECTION BASED ON INFORMATION THEORY*. Department of Computer Science and Articial Intelligence, UNIVERSITY OF ALICANTE
- 2010. p.
- BROWN, G. *A New Perspective for Information Theoretic Feature Selection*. En: Proceedings of 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA, 2009. p.
- CHENG, H.; Z. QIN, *et al.* Conditional Mutual Information-Based Feature Selection Analyzing for Synergy and Redundancy *ETRI*, 2011, 33(2): 210-219.

- DING, C. and H. PENG Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Bioinformatics and computational biology*, 2005, 3(2): 185-205.
- FLEURET, F. Fast Binary Feature Selection with Conditional Mutual Information. *Machine Learning Research*, 2004, 5: 1531-1555.
- FREY, L. and D. FISHER. *Identifying Markov Blankets with Decision Tree Induction*. Third IEEE International Conference on Data Mining, 2003. p.
- GUYON and A. ELISSEEFF An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003, 3: 1157-1182.
- HALL, M. A. *Correlation-based Feature Selection for Machine Learning*. Department of Computer Science. Hamilton, New Zealand, University of Waikato, 1999. 149. p.
- JOHN, G. H.; R. KOHAVI, *et al. Irrelevant features and the subset selection problem.* International Conference in Machine Learning, 1994. 121-129 p.
- KHINCHIN, A. I. *Mathematical foundations of information theory*. 1957. p.
- KOHAVI, R. and G. H. JOHN Wrappers for feature subset selection. *Artif. Intell.*, 1997, 97(1–2): 273-324.
- KOLLER, D. and M. SAHAMI. *Toward optimal feature selection.* Thirteenth International Conference on Machine Learning, Bari, Italia, 1996. 284-292. p.
- LI, W. and Y. YANG How many genes are needed for a discriminant microarray data analysis? *Methods of microarray data analysis*, 2002: 137-150.
- MEYER, P.; C. SCHRETTER, *et al.* Information-theoretic feature selection in micro-array data using variable complementarity *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, 2008, 2: 261-274.
- MEYERS, L. S.; G. GAMST, *et al. Applied Multivariate Research: Design and Interpretation*. SAGE Publications, 2005. p. 9781412904124
- SAEYS, Y.; I. INZA, *et al.* A review of feature selection techniques in bioinformatics *Bioinformatics and computational biology*, 2007, 23(19): 2507–2517.
- SHANNON, C. and W. WEAVER. *The Mathematical Theory of Communication*. Urbana, IL, University of Illinois Press, 1963. p.
- XIONG, M.; Z. FANG, *et al.* Biomarker identification by feature wrappers. *Genome Research*, 2001, 11: 1878-1887.

- YANG, J. and J. O. PEDERSEN. *A Comparative Study on Feature Selection in Text Categorization*. Fourteenth International Conference on Machine Learning (ICML'97), 1997. 412-420 p.
- YEUNG, K. Y. and R. E. BUMGARNER Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biology*, 2003, 4(12): R83.
- YU, L. and H. LEI Efficient Feature Selection via Analysis of Relevance and Redundancy *Machine Learning Research*, 2004, 5: 1205-1224.
- ZHENG, Y. and C. K. KWOH A feature subset selection method based on high-dimensional mutual information *Entropy*, 2011, 13: 860-901.