Selección estable de variables independientes con RFE

Mauro Di Masso y Pablo M. Granitto

CIFASIS
Centro Internacional Franco Argentino de Ciencias de la Información y Sistemas
UPM (Francia) / UNR–CONICET (Argentina)
Bv 27 de Febrero 210 Bis, 2000 Rosario, República Argentina
{granitto}@cifasis-conicet.gov.ar

Abstract. La selección de variables es un problema de interés práctico actual en el campo de los sistemas inteligentes, con numerosas e importantes aplicaciones. Uno de los métodos más exitosos es el de Recursive Feature Elimination o RFE. Este método presenta dificultades cuando se analizan problemas con variables altamente redundantes. En este trabajo exploratorio se introduce una variante RFE que busca solucionar este inconveniente. El nuevo SRFE incorpora un vector de penalización tal que, dado un grupo de variables redundantes, una sola de ellas no es penalizada y todas las demás sí lo son. La forma de elegir que variable se penaliza incorpora un simple criterio para facilitar la estabilidad de las selecciones. El nuevo método se compara con otros desarrollos similares en dos datasets, mostrando evidencias que SRFE elige conjuntos de mejor calidad, tanto en los niveles de error conseguidos, la selección de variables independientes, como en la estabilidad de la solución.

1 Introducción

Muchos problemas de interés actual en Machine Learning comparten la característica de tener como entrada muchas variables, a veces muchas más que el número de ejemplos disponibles (datasets anchos). Entre los ejemplos más importantes se cuentan la química in-silico [1], las tecnologías de "high-throughput" en biología [2] o la comprensión de textos. En estos casos, usualmente la mayoría de las variables medidas tienen una importancia relativamente baja para resolver el problema en cuestión [3], y algunas veces llegan a interferir con el aprendizaje en lugar de ayudarlo, fenómeno que se suele llamar "la maldición de la dimensionalidad".

La selección de variables es una técnica de pre-procesado que ayuda a resolver este problema [4]. El principal objetico es encontrar un conjunto reducido de las variables originales que mejore, o al menos no empeore, la performance del método de modelado aplicado al dataset en estudio. La selección de variables no solo alivia el problema de la dimensionalidad, también permite una simplificación de los modelos utilizados, una visualización más simple de los datos y en general

una mejor interpretación tanto de los datos como de los modelos desarrollados [5].

En este contexto, el algoritmo "Recursive Feature Elimination" (RFE) muestra muy buenos resultados con una carga computacional moderada, en especial para problemas anchos [6]. La versión original y más utilizada de este método usa una Support Vector Machine [7] con kernel lineal para ordenar las variables, aunque hay versiones que utilizan otros métodos de modelado en reemplazo de la SVM [8].

Hay dos situaciones por las cuales una variable debería ser removida o eliminada durante el proceso de selección. La primera y más simple de entender es cuando una variable tiene poco o nula información sobre la propiedad en estudio. Los wrappers en general [4], y RFE en particular, son eficientes para resolver esta situación. La segunda es cuando un par (o un grupo) de variables aportan la misma información sobre el problema, porque tienen una alta correlación entre sí (sea correlación lineal como no-lineal). RFE tiene un problema conocido ante esta segunda situación, ya que tiende a dar igual importancia a todas las variables del grupo en las etapas iniciales. En algunos casos el método logra concentrar la importancia en alguna de las variables del grupo, aunque sin un criterio fijo, lo que produce un problema de estabilidad ante experimentos repetidos con pequeñas diferencias entre los mismos. Varios trabajos previos, que se discuten en la próxima sección, propusieron soluciones parciales a este problema de correlación durante la selección. Nosotros proponemos en este trabajo un nuevo método, "Stable Recursive Feature Elimination" (SRFE) que apunta a resolver al mismo tiempo los dos problemas mencionados, la seleccion de variables correlacionadas y la estabilidad de las soluciones.

2 Trabajos previos

En un importante trabajo en selección de variables, donde analizan las nociones de importancia y redundancia, Yu y Liu [9] propusieron seleccionar variables usando un filtro que combine una medición de la importancia de la variable con una penalización por la redundancia de la misma, el "Fast Correlation Based Filter" (FCBF). El método propone en primer instancia medir la información mutua entre cada variable y la clase, utilizando una medida normalizada, la symmetrical uncertainty (SU) [9], y descartar -poner al final del ranking- todas las variables con SU menor que un umbral dado. En un segundo paso, el método propone medir la SU de cada variable con todas las demás, y eliminar de forma recursiva todas las variables que tienen alta SU con alguna de las variables ya seleccionadas. La ventaja de este método es su velocidad, ya que no requiere ajustar clasificadores, pero su eficacia es muy limitada ante variables que funcionan en conjunto, como en el caso del conocido problema del XOR de dos variables.

El método RFE es un simple proceso recursivo que ordena las variables de acuerdo a una medida de la importancia de cada variable dada por un clasificador [6]. A cada iteración se mide la relevancia de todas las variables y la menos im-

portante es removida. En la práctica, para acelerar el proceso, en cada iteración se remueve un grupo de variables, típicamente un porcentaje bajo del total de variables presentes. La recursión en el ordenado de las variables permite mejorar la performance cuando hay variables correlacionadas. En la versión original la importancia está dada por la componente del vector perpendicular al plano de separación entre las clases en la dirección de cada variable. Ante la presencia de un par de variables similares las SVM reparten la importancia en partes iguales entre las mismas. Recién cuando el método logra eliminar una de estas variables la otra toma toda la importancia correspondiente, y al iterar el ordenamiento esta variable suele mejorar notoriamente su posicion en el ranking [10]. Esta solución parcial a la dificultad descripta sufre de un claro problema de estabilidad. Cuál variable es removida y cuál queda es una decisión casi al azar, tomada en base a la pequeña diferencia en importancia que puede haber entre ellas, lo que produce que en los típicos experimentos replicados sobre datos anchos ambas variables aparezcan seleccionadas en la mitad de los experimentos y no figuren en la otra mitad, lo que dificulta la interpretación de los resultados enormemente.

Mundra et al. [11] presentaron recientemente el método conocido como "minimum-redundancy maximum-relevancy" (MRMR). La propuesta consiste en agregar al orden creado por RFE una penalización que considere la redundancia entre variables. Para la i-ésima variable se define

$$r_i = \beta|w_i| + (1-\beta)\frac{R_i}{Q_{S,i}},$$

donde el primer término es el original de RFE y el segundo es la nueva penalización. Este término es un cociente entre una medida de relevancia de la variable ante las clases (R_i) y una de redundancia de la variable ante las demás variables $Q_{S,i}$ [11]. $\beta \in [0,1]$ es un parámetro que define el peso relativo de cada término. La solución planteada evita la presencia de variables redundantes, pero no es la adecuada. Cuando se encuentra un par (o un grupo) de variables redundantes se penaliza a todas por igual en lugar de determinar un criterio por el cual una de ellas es mejor que las otras. En esta situación una variable poco relevante y poco redundante puede llegar a ocupar un lugar alto en el ranking final, desplazando variables que sí son relevantes al problema pero que están fuertemente penalizadas.

3 RFE con Penalización de redundancia

Para evitar el problema de la redundancia, el algoritmo SRFE que proponemos aquí cuenta en primera instancia con un filtro inspirado en FCBF, encargado de determinar los niveles de correlación y decidir qué variables aportan similar información en el momento del aprendizaje. Con este filtro se genera un vector de penalizaciones $\mathbf{P}$ que actúa sobre todas menos una de las variables redundantes, siguiendo además un criterio que considera la estabilidad de la solución. $\mathbf{P}$ luego se combina con el ranking RFE, siguiendo la idea del método MRMR, para determinar qué variables eliminar.

De acuerdo a este diseño la relevancia de la variable i en el problema queda determinada por:

$$r_i = \beta|w_i|_e + (1-\beta)P_i,$$

similar a la utilizada en MRMR, con la única diferencia en la forma de penalizar a las variables redundantes dada por P_i. El subíndice e en el primer término indica que el vector de pesos de la SVM se escala para que tenga el mismo rango que los P_i.

El vector de penalización $\mathbf{P}$ se calcula como paso inicial del método. Siguiendo a FCBF, usamos la SU entre pares de variables para medir la redundancia (SU_{ij}), y la SU entre variables y la clase para medir la importancia de cada una ($SU_{i,c}$). La construcción es iterativa, comenzando con todas las variables. En cada paso se busca el par de variables i, j con mayor SU entre ellas. SRFE, a diferencia de MRMR, penaliza a una sola de las variables del par, que es la que tiene menor información de la clase (por ejemplo, i en este caso) con el valor $P_i = -SU_{ij}$. Si los valores de $SU_{i,c}$ y $SU_{j,c}$ difieren en menos de un umbral de tolerancia T_p entonces se considera que las variables i y j aportan el mismo valor predictivo al modelo y, para mantener un criterio estable, simplemente se penaliza a aquella variable con mayor subíndice. La variable penalizada se remueve del conjunto, y se itera el procedimiento hasta completar el vector $\mathbf{P}$. La severidad de la penalización queda dada por el nivel de redundancia mismo entre las variables. Durante la experimentación, el umbral T_p fue establecido en 0.05, un valor bajo, para evitar solamente perturbaciones dadas por el ruido y el muestreo aleatorio. Si bien para n variables el procedimiento tiene complejidad $O(n^2)$, el cálculo es sencillo y se realiza una sola vez, amortizándose frente a las n iteraciones posteriores del clasificador. El vector $\mathbf{P}$ resultante tiene una sola variable con penalización 0 y las demás tienen un valor $P_i \in [-1, 0]$, dependiendo del nivel de redundancia medido. Esto se debe a que al usar muestras finitas $SU_{ij} > 0$ siempre, aún para variables independientes. Pera evitar este problema numérico se eliminan en un último paso todas las penalizaciones $P_i < T_c$. El umbral T_c es el valor de SU para el cual dos variables ya no se consideran redundantes, y fue fijado en 0.1 en este trabajo.

Una vez obtenido $\mathbf{P}$ se utiliza el procedimiento general de RFE, recalculando a cada paso solamente la SVM correspondiente.

4 Experimentos

4.1 Disposición Experimental

Usamos una disposición adecuada para evaluar selección de variables [8]. El proceso consiste a cada paso de dos bucles anidados. El externo repite 50 veces una división al azar del dataset en un conjunto de entrenamiento con el 75% de los datos (usado para entrenar los modelos y elegir las variables) y un conjunto de test con el resto, usado solamente para medir el error de clasificación del modelo generado. El bucle interno realiza la selección propiamente dicha y el ajuste de los correspondientes clasificadores, incluyendo la selección de parametros como

la constante C de las SVM con una validación cruzada que sólo usa esos datos de ajuste. Se tomó un valor fijo de $\beta = 0.5$ para MRMR y SRFE. El resultado de las 50 réplicas del experimento se promedian para obtener curvas de error para cada método de selección, y las variables elegidas se analizan para ver su estabilidad ante las pequeñas variaciones en los datos de entrenamiento. Se comparan los 4 métodos de selección descriptos en el trabajo: FCBF, MRMR, RFE y SRFE, en todos los casos usando el mismo clasificador (SVM) y los mismos conjuntos de ajuste y test.

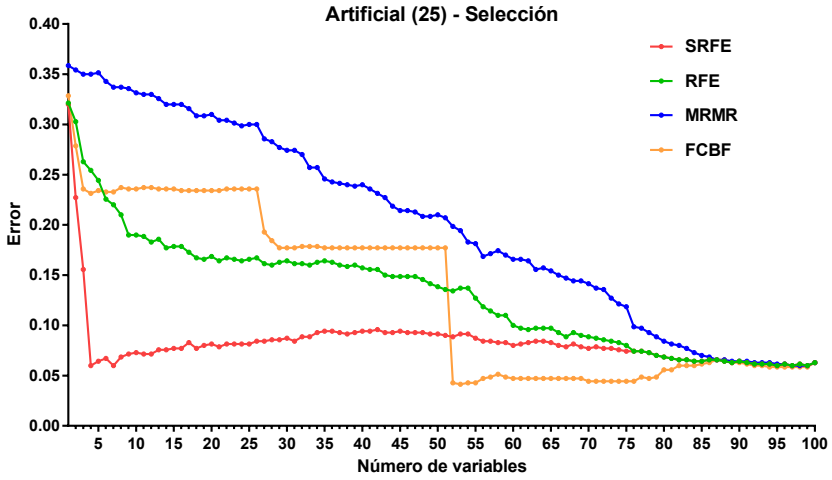
En este trabajo analizamos resultados sobre dos datasets. En primer lugar usamos un dataset artificial diseñado para resaltar los problemas de redundancia y estabilidad. El dataset "artificial" consta de 100 variables, agrupadas en 4 conjuntos altamamente correlacionados de 25 variables. A su vez, el concepto objetivo, que es binario, está determinado por una conjunción de los 4 grupos, lo que significa que es necesaria la presencia de al menos una variable de cada grupo para obtener la solución correcta. El dataset tiene 25 puntos. En segundo lugar usamos el dataset "fragola", con mediciones reales sobre muestras de 9 distintas especies de frutillas realizadas con un espectrómetro de masa de reacción de transferencia de protón (PTR-MS) acoplado con un detector de tipo cuadrupolo. El espectrómetro mide la composición del aire alrededor de cada frutilla (el "olor") y da como respuesta la proporción de compuestos para cada masa atómica desde 1 hasta 250 AMU, aproximadamente. El dataset contiene 231 registros medidos sobre 233 variables. La espectrometría PTR-MS presenta típicamente altas correlaciones entre grupos de pocas variables, muchas veces 2 o 3, debido a la existencia de isótopos y fragmentos de moléculas.

4.2 Resultados

En la Figura 1 se comparan las tasas de error para los cuatro métodos sobre el dataset Artificial. El filtro FCBF no tiene la capacidad de detectar la interacción entre las variables de los distintos grupos, por lo que considera a todas las variables como equivalentes, y las elimina de forma ordenada. Cada salto en la curva de error corresponde a la eliminación de la última variable de cada grupo. RFE logra ver las interacciones, al estar basado en un modelo multivariado, pero sufre el problema descripto en la discusión previa ante la gran redundancia de variables. MRMR es en este caso la peor solución, al penalizar en la practica a todas las variables por igual. SRFE obtiene la solución correcta, con un claro mínimo global en el error medio con 4 variables. En el caso real del dataset Fragola (Figura 2) las diferencias entre los métodos para los niveles de error son mucho menores, pero igualmente SRFE muestra el menor error medio para un número bajo de variables.

Las diferencias entre los métodos son más notables al analizar además la calidad y estabilidad de las soluciones. En la Figura 3 se muestran los resultados correspondientes. En la columna de la izquierda se muestra la proporción de veces que cada variable fue seleccionada entre las 4 más importantes por cada algoritmo para el problema artificial. Solamente SRFE fue capaz de seleccionar de manera estable las mismas 4 variables en casi todas las corridas, 1 variable de

Fig. 1. Error medio de clasificación como función del número de variables seleccionadas por los 4 métodos evaluados en el trabajo para el dataset Artificial.



cada grupo. Los otros 3 métodos seleccionaron variables redundantes y de forma inestable. En la columna de la derecha se muestran los resultados correspondientes para el dataset Fragola, pero tomando para el análisis las mejores 6 variables, que es el mínimo error con SRFE. En este caso MRMR y SRFE muestran ambos las soluciones más estables. Sin embargo, MRMR selecciona variables con alta correlación correspondientes a isótopos de los mismos compuestos (masas 103-104 y 131-132), mientras que SRFE consigue siempre desarmar estos pares de variables relevantes pero redundantes.

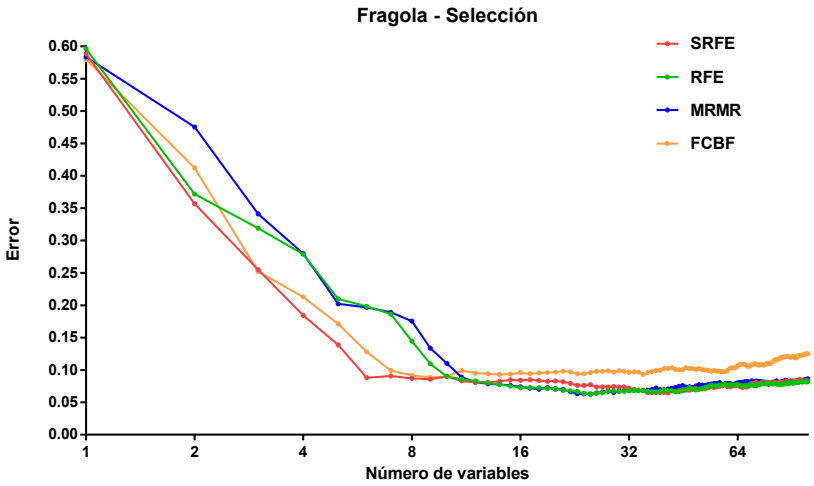
5 Conclusiones

En este trabajo exploratorio se introdujo una variante del conocido método de selección de variables RFE. SRFE incorpora un vector de penalización por redundancia construído de forma tal que, dado un grupo de variables redundantes, una sola de ellas no es penalizada y todas las demás sí lo son. La forma de elegir que variable se penaliza incorpora un simple criterio para facilitar la estabilidad de las selecciones.

El nuevo método se comparó con otros desarrollos similares en selección de variables. Usando un dataset artificial y otro real, en ambos casos con gran cantidad de variables y pocos ejemplos, se mostró que SRFE elige conjuntos de mejor calidad que los otros métodos, tanto en los niveles de error conseguidos, la selección de variables independientes, como en la estabilidad de la solución.

En un trabajo próximo se discutirá la dependencia del método con los valores que toman los dos parámetros libres T_c y T_p y se presentarán resultados sobre numerosos datasets.

Fig. 2. Error medio de clasificación como función del número de variables seleccionadas (en escala logarítmica) por los 4 métodos evaluados en el trabajo para el dataset Fragola.



Agradecimientos

PMG agradece a la ANPCyT por financiamiento parcial (proyecto PICT 2012-0181).

References

1. H. Li, C. Ung, C. Yap, Y. Xue, Z. Li, Z. Cao, Y. Chen, Prediction of genotoxicity of chemical compounds by statistical learning methods, Chemical research in toxicology 18 (6) (2005) 1071–1080.
2. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science 286 (5439) (1999) 531–537.
3. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.
4. R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1) (1997) 273–324.
5. H. Liu, E. R. Dougherty, J. G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, et al., Evolving feature selection, Intelligent systems, IEEE 20 (6) (2005) 64–76.
6. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1-3) (2002) 389–422.
7. V. Vapnik, The nature of statistical learning theory, springer, 2000.
8. P. M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products, Chemometrics and Intelligent Laboratory Systems 83 (2) (2006) 83–90.

9. L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, The Journal of Machine Learning Research 5 (2004) 1205–1224.
10. C. Furlanello, M. Serafini, S. Merler, G. Jurman, Entropy-based gene ranking without selection bias for the predictive classification of microarray data, BMC bioinformatics 4 (1) (2003) 54.
11. P. A. Mundra, J. C. Rajapakse, Svm-rfe with mrmr filter for gene selection, NanoBioscience, IEEE Transactions on 9 (1) (2010) 31–37.

Fig. 3. Estabilidad de las variables seleccionadas por los diferentes métodos para el dataset Artificial (izquierda) y Fragola (derecha).