A new feature subset selection using bottom up clustering

David Quesada López

Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain

Objetivos

Queremos seleccionar un subconjunto de variables con un método de filtrado basado en cluster jerárquico para:

- Reducir la dimensionalidad de nuestro dataset.
- Conseguir un subconjunto de variables relevantes y no redundantes que sean igual o más eficientes y eficaces que el conjunto inicial.
- Aprovechar tanto la velocidad del filtrado como la capacidad para agrupar variables del clustering.

Introducción

La selección del subconjunto de variables es una parte esencial en el preprocesado de los datos para reducir la dimensionalidad. Existen vários métodos para esto:

- Wrapper [2], que seleccionan subconjuntos de variables en base a su precisión pero son muy costosos computacionalmente.
- Filter [3], que miden la relevancia de las variables sin mucho coste pero sin asegurar buena precisión final.

En nuestro algoritmo CFSS usaremos un método de filtrado basado en clustering jerárquico para agrupar las variables redundantes en clusters y elegir las relevantes para nuestro subconjunto.

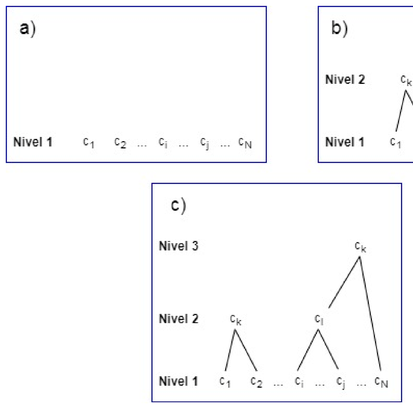

Figure 1: Cluster *bottom-up* de variables

Materiales

Para completar esta investigación se han usado los siguientes materiales:

- 11 datasets válidos del repositorio UCI para probar el funcionamiento de CFSS.
- Los métodos mRMR, ReliefF y L1-LSMI para comparar el rendimiento de CFSS.
- El algoritmo GACH para acotar el número de variables del subconjunto final.
- Un clasificador kNN en el que probar los subconjuntos devueltos por los métodos anteriores en sus respectivos datasets.

Unión de CFSS y GACH

CFSS agrupa las variables en clusters para reducir la redundancia y elegir las variables por ranking y GACH determina un valor adecuado para el tamaño del subconjunto.

Agrupación de variables

Para agrupar las variables en clusters se sigue un criterio de similaridad entre ellas, en vez de distancia, basado en la información mutua [1]:

$$I(f_i, f_j) = \sum\sum p(f_i, f_j) log \frac{p(f_i, f_j)}{p(f_i)p(f_j)} \qquad (1)$$

Siendo $p(f_i)$ la función de distribución de esa variable. Dentro de un cluster, el centroide será la variable con una mayor información mutua con la clase. Este centroide determina la relevancia de un cluster.

Se agruparán variables que no quede ningún cluster con un sólo elemento. En ese punto, se eligen qué variables se meten al subconjunto final: se empieza a añadir desde los clusters más relevantes. De un mismo cluster se pueden meter más de una, y de clusters menos relevantes que un mínimo establecido se pueden ignorar sus variables.

Número de variables

Una vez que hemos agrupado las variables con CFSS, utilizamos el algoritmo GACH para establecer un tamaño adecuado del subconjunto de variables. GACH sigue también un cluster *bottom-up*, fusionando clusters hasta que sólo quede uno. Tras esto, devuelve los p-valores obtenidos en cada paso de fusión. Los puntos con saltos considerables en los p-valores son candidatos a ser elegidos como punto de corte.

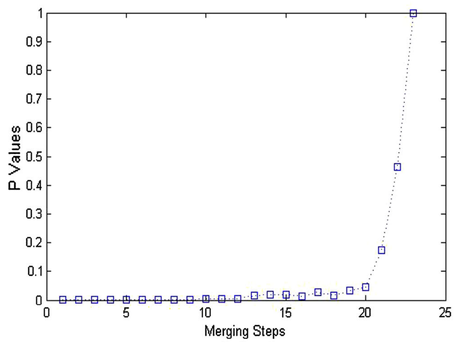

Figure 2: Estimación del número de variables con GACH

Resultados

CFSS obtiene mejores resultados en el promedio de todos los datasets que el resto de algoritmos de selección de variables probados.

Promedios	ReliefF	mRMR	L1-LSMI	CFSS
Tiempo (seg.)	3.17	12.84	28.09	**2.63**
Precisión	80.41	81.25	77.68	**83.17**

Table 1: Resultados promedio

Conclusión

El método CFSS basado en clustering jerárquico propuesto representa una alternativa nueva a la hora de seleccionar un subconjunto de variables. Este agrupa las variables redundantes en clusters y elige sólo variables relevantes para el subconjunto final. El algoritmo GACH proporciona una buena estimación del número de variables que se deberían elegir. Los resultados de CFSS obtenidos son buenos en comparación con los algoritmos alternativos propuestos. En trabajo futuro habría que estudiar mejor cuál es el número de variables representativas de cada cluster y una condición de parada más robusta a la hora de fusionar clusters.

Artículo original

Dehghan Z., Mansoori E. G. (2016). A new feature subset selection using bottom-up clustering. *Pattern Analysis and Applications*, 1-10.

Referencias

[1] Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

[2] Kohavi R, John GH (1997) Wrapper for feature subset selection. Artif Intell 97(1–2):273–324

[3] Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundance. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

POLITÉCNICA