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 [4], 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.

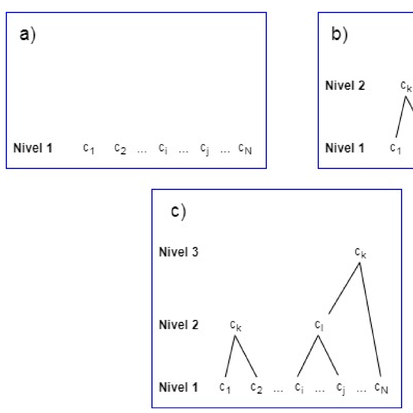

Figure 1: Cluster *bottom-up* de variables

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 clusters de CFSS.
- Un clasificador kNN en el que probar los subconjuntos devueltos por los métodos anteriores en sus respectivos datasets.

Important Result

Lorem ipsum dolor **sit amet**, consectetur adipiscing elit. Sed commodo molestie porta. Sed ultrices scelerisque sapien ac commodo. Donec ut volutpat elit.

Agrupación de variables

Para agrupar las variables en clusters 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 agrupan variables en clusters hasta que no queda 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.

Métodos

El algoritmo comienza asignando cada variable a un cluster diferente (*bottom-up*). Para medir la distancia entre dos variables miramos su similitud en base a su información mutua:

Results

Placeholder

Image

Figure 2: Figure caption

Nunc tempus venenatis facilisis. Curabitur suscipit consequat eros non porttitor. Sed a massa dolor, id ornare enim:

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table 1: Table caption

Conclusion

Nunc tempus venenatis facilisis. **Curabitur suscipit** consequat eros non porttitor. Sed a massa dolor, id ornare enim. Fusce quis massa dictum tortor **tincidunt mattis**. Donec quam est, lobortis quis pretium at, laoreet scelerisque lacus. Nam quis odio enim, in molestie libero. Vivamus cursus mi at *nulla elementum sollicitudin*.

Artículo real

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