Selección del subconjunto de variables y clustering

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

Abstract

La selección del subconjunto de variables es el proceso de encontrar un subconjunto de las variables originales que genere los mejores resultados comparado con todo el conjunto de variables en términos de la eficiencia y la efectividad. Los métodos de selección del subconjunto de variables están divididos en tres categorías dependiendo del criterio de evaluación: métodos filter, wrapper o embebidos. En este estado del arte, revisaremos por encima estas tres variantes y nos centraremos en los métodos de filtrado que utilizan clustering.

KEY WORDS: Feature subset selection; Filter method; Feature clustering;

1 Introducción

Cuando te enfrentas a un problema de aprendizaje automático, uno de los inconvenientes que puedes tener es un gran número de variables en tu dataset. De estas variables no todas tienen por qué ser relevantes para tu problema, y en algunos casos usar todo el conjunto puede disminuir la efectividad de nuestro sistema (H. Liu X. *et al.* (2011)).

El problema de elegir el subconjunto de variables óptimo es un problema NP-completo (A.A. Albrecht (1982)), por lo que el objetivo es encontrar el subconjunto de las variables que nos satisfaga en términos de efectividad y eficiencia, no el óptimo. Cuando hablamos de efectividad nos referimos a la precisión final de nuestro sistema que obtenemos a partir del subconjunto de las variables, y al hablar de eficiencia estamos hablando sobre el tiempo que tardamos en obtener dicho subconjunto.

Para atacar este problema se distinguen tres puntos de vista diferentes en base al criterio seguido para obtener el subconjunto de variables útiles:

- Si la selección de este subconjunto está basada en la precisión predictiva de nuestro algoritmo de aprendizaje
- filter
- embeded ...

2 State-of-the-art

Las variables que buscamos obtener en nuestro subconjunto deben ser relevantes y no redundantes entre sí. Dependiendo de si un algoritmo distingue por un tipo o por ambos aparecen distintos tipos de filtering.

2.1 More specific

Perhaps some subsections are needed.

3 Conclusions and future research

What are the main open lines for research.

References

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*. 2nd edn, J. Wiley and Sons, New York.

Shakhnarovich, G., El-Yaniv, R. and Baram, Y. (2001) Smoothed bootstrap and statistical data cloning for classifier evaluation. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, Morgan Kaufmann, pp. 521–528.

Uncu, O. and Türksen, I.B. (2007) A novel feature selection approach: Combining feature wrappers and filters. *Information Sciences*, 177, 449–466.

H. Liu, X. Wu, and S. Zhang, "Feature selection using hierarchical feature clustering", in *Proc. ACM Int. Conf. Inform. Knowl. Manage.*, New York, NY, USA, 2011.

Wold, H. (1975) Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In Gani, J. (ed), *Perspectives in Probability and Statistics*, Academic Press, London, pp. 117–142.

A.A. Albrecht, Stochastic local search for the feature set problem, with applications to microarray data, Applied Mathematics and Computation 183 (2006) 1148–1164