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, haciendo especial mención de aquellos 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 resultante, entonces nos encontramos ante un método wrapper. En este caso, las variables van siendo tenidas en cuenta en el modelo si mejoran su capacidad predictiva. El modelo es usado como una caja negra para ver su capacidad predictiva. Los algoritmos wrapper (Kohavi and John (1997)) son bastante eficaces en cuanto a precisión del modelo resultante, sin embargo sufren de una carga computacional muy elevada que los hace inabarcables en datasets con un conjunto muy grande de variables.

- Si la selección del subconjunto es independiente del algoritmo de aprendizaje usado después, entonces hablamos de un método de filtrado o filter. Estos métodos se basan en medidas matemáticas como la información mutua o el coeficiente de correlación de Pearson para relacionar unas variables con otras y medir su relevancia (Gheyas and Smith (2009)). Estos métodos no son muy costosos computacionalmente comparados con los métodos wrapper y no están ligados al algoritmo de aprendizaje usado después, sin embargo no garantizan la precisión del modelo final (Song *el al.* (2013)).

- Si se usa la selección de variables como parte del algoritmo de aprendizaje, entonces se habla de métodos embebidos o embedded. Ejemplos clásicos de estos métodos son los árboles de decisión o las redes neuronales (Breiman *el al.* (1984)).

- También existen los llamados métodos híbridos, que combinan la rapidez de los métodos de filtrado con los buenos resultados de los métodos wrapper (Uncu and Türksen (2007)).

Este problema es también dividido a veces en supervisado, si tenemos algunas etiquetas que indiquen qué variables son útiles, o no supervisado, si no tenemos ninguna indicación de la utilidad de las variables (Li *el al.* (2014)).

En este estado del arte nos centraremos en los métodos de filtrado. El resto del documento se organiza en la sección 2, donde se repasan distintos métodos de filtrado y las técnicas que se utilizan, y la sección 3, donde se habla de las líneas de investigación abiertas en este ámbito.

2 State-of-the-art

Al elegir un subconjunto de variables, tenemos que fijarnos en que estas sean relevantes y no redundantes entre sí. Una variable es relevante cuando ayuda a discernir más la clase de una instancia que otra variable, y es redundate cuando aporta la misma información sobre la clase a la que se pertenece que otra variable. (Song *el al.* (2013)) Que una variable no sea relevante provoca empeoramientos en la precisión de nuestro algoritmo de aprendizaje (John Kohavi and Pfleger (1994)), y la redundancia entre variables provoca deterioros en el rendimiento de nuestro modelo, por lo que conviene evitar ambos casos.

Los primeros métodos que aparecieron para el filtrado se centraban en encontrar variables relevantes para el modelo, uno de los ejemplos más clásicos es el algoritmo *Relief* (Kira and Rendell (1992)). *Relief* discrimina unas variables de otras dependiendo de lo bien que diferencian la clase según un criterio basado en análisis estadístico de la distancia. También aparece el concepto de Markov blanket de las variables para medir su relevancia en presencia de otras variables (Mitra *et. al* (2002)).

Como primera aproximación, es un buen modo de reducir en parte la dimensionalidad del conjunto de variables, pero este algoritmo no comprueba que las variables que selecciona estén muy correlacionadas entre sí, por lo que puede seleccionar variables redundantes entre sí que no ayuden a mejorar la precisión del modelo y sin embargo dejar fuera variables que por sí mismas no están muy correlacionadas con la clase pero que aportan más información en conjunto con las que tenemos.

Para solventar el problema de la redundancia entre variables se crean nuevos algoritmos como *CFS* y *FCBF* que emplean medidas como la información mutua y la incertidumbre simétrica para comprobar la redundancia entre variables (Yu and Liu (2004)). Estos métodos son bien conocidos y referenciados a la hora de presentar un nuevo algoritmo y probar su rendimiento.

En este punto donde ya se tienen medidas para calcular la relevancia y la redundancia de una variable aparecen nuevas formas no supervisadas de aplicar esta medida por medio del clustering de variables.

2.1 Filtrado por clustering

Los primeros problemas a los que se intentó aplicar esta aproximación son la clasificación de textos por medio de clustering jerárquico (Dhillon *et. al* (2003)) y los microarrays de ADN (Yu and Liu (2004)). En ambos casos, la dimensionalidad de las variables es muy alta y en el caso de los microarrays no se tienen muchas instancias del problema.

A raíz de los buenos resultados obtenidos, se continúa investigando la rama del clustering en sus diferentes variantes como solución a este problema. Hay investigadores que se centran en el clustering jerárquico o aglomerativo clásico y en encontrar medidas más efectivas para la distancia entre clusters

de variables y la forma de unir estos, como el el caso del algoritmo *FSFC* (H. Liu X. *et al.* (2011)) y más recientemente el algoritmo de Dehghan and Mansoori (2016).

Hay otros investigadores que prefieren buscar resultados mejores en el clustering particional en vez del jerárquico y tratan de encontrar medidas adecuadas para ver la similitud de las variables que se encuentran en un mismo cluster o partición, como en el caso de *RPCCL* (Yiu and Hong (2012)).

Otra rama de investigación se centra en nuevos métodos para realizar el cluster de variables, como en el caso del algoritmo *FAST* (Song *el al.* (2013)) donde usan un *minimum-spannin tree* para hacer el clustering y seleccionar el subconjunto de variables particionando este árbol.

3 Conclusiones y ramas de investigación abiertas

El uso del clustering para la selección del subconjunto de variables es un método que resulta muy eficaz para seleccionar variables relevantes y no redundantes. Aplicar clustering evita el sobreajuste y como método de filtrado es bastante eficiente.

En este momento, la investigación en este área se centra en la obtención de nuevas medidas de la distancia basadas en la redundancia y la relevancia de las variables, así como en la obtención de nuevos métodos de clustering de variables.

Sería interesante intentar aplicar el clustering probabilístico a este problema, dado que los otros dos tipos de clustering clásicos ya han sido probados. También parece ser una buena rama el uso de métodos híbridos de filtrado y de wrapper, por lo que aplicar en conjunto alguno de los algoritmos de clustering vistos con métodos wrapper podría obtener bastantes buenos resultados.

References

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.

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

R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997.

I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognition (2009), doi: 10.1016/j.patcog.2009.06.009.

Qinbao Song, Jingjie Ni and Guangtao Wang, A fast clustering-based feature subset selection algorithm for high dimensional data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1* year 2013.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth and Brooks, 1984.

Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou,and Hanqing Lu, Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:26 NO:9* September 2014.

John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

Kenji Kira, Larry A. Rendell. A practical approach to feature selection, 1992.

L. Yu, H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5: 1205-1224, 2004.

Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3,pp 1265-1287, 2003.

Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray data, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2001.

P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):301–312, 2002

Yiu-ming Cheung and Hong Jia, Unsupervised Feature Selection with Feature Clustering. *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*

Dehghan Z. , Mansoori E.G., A new feature subset selection using bottom-up clustering. Pattern Anal Applic (2016). https://doi.org/10.1007/s10044-016-0565-8