Atipicidad: Medida de calidad clave dentro del descubrimiento de reglas descriptivas supervisadas

C.J. Carmona[1], M.J. del Jesus[1], F. Herrera[2]
[1]*Departamento de Informática. Universidad de Jaén, España* ccarmona@ujaen.es, mjjesus@ujaen.es
[2]*Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada, España* herrera@decsai.ugr.es
[1,2]*Instituto Data Science and Computational Intelligence (DaSCI)*

***Resumen*—Esto es un resumen de nuestro trabajo publicado en la revista *Knowledge-Based Systems* [1] para su presentación en la Multiconferencia CAEPIA'18 Key Works.**

***Index Terms*—Descubrimiento de reglas descriptivas supervisadas, Descubrimiento de subgrupos, Conjuntos de contraste, Patrones emergentes, Atipicidad.**

I. Resumen

A lo largo de la literatura podemos encontrar un conjunto de técnicas que se encuentran a medio camino entre la predicción y la descripción, agrupadas bajo el nombre de descubrimiento de reglas descriptivas bajo aprendizaje supervisado (SDRD) [1], [2]. Este conjunto de técnicas intenta obtener reglas de una categoría o clase prefijada para describir información significativa y relevante del conjunto de datos.

El principal objetivo de las técnicas dentro de SDRD no es extraer un modelo para clasificar nuevas instancias, sino obtener un modelo que permita comprender, describir o encontrar fenómenos subyacentes de interés en los datos. Dentro de este área de investigación se agrupan todas aquellas técnicas que, mediante el uso de reglas y un modelo de aprendizaje supervisado, intentan obtener conocimiento descriptivo de los datos que sea significativo, inusual y de interés para el usuario, como el descubrimiento de subgrupos [3], la minería de patrones emergentes [4] y los conjuntos de contraste [5], entre otros. A continuación se describen brevemente estas técnicas.

Descubrimiento de subgrupos: Se define como el descubrimiento de subgrupos de la población estadísticamente interesantes, esto es, tan grandes como sea posible y con una distribución estadística de la propiedad de interés lo más atípica posible respecto al conjunto de la población. La medida de calidad para medir esta atipicidad [6] en una regla R se define como:

$$WRAcc(R) = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right) \quad (1)$$

Conjuntos de contraste: Es una técnica de exploración para contrastar diferencias de grupos dentro de un conjunto de datos, es decir, descubrir conjuntos de ejemplos con amplias diferencias de soporte entre grupos del conjunto de datos y se mide mediante la diferencia de soporte [5]:

$$DS(R) = \left|\frac{p}{P} - \frac{n}{N}\right| \ \geq \ \delta \quad (2)$$

Patrones emergentes: Se centra en buscar conocimiento relacionado con los valores de una clase, donde el número de instancias cubiertas por un patrón sea muy elevado para un valor de la variable objetivo y muy bajo o nulo para el resto; es decir, que el mismo patrón tenga un soporte muy alto para una clase y muy bajo para las demás clases del problema, y viene dado por el índice de crecimiento [4] que se representa a continuación:

$$GR(R) = \frac{\frac{p}{P}}{\frac{n}{N}} > 1 \quad (3)$$

Centrándonos en la importancia de la medida de atipicidad hay que destacar que mide el equilibrio entre generalidad y precisión. En concreto, el segundo factor de la atipicidad es el factor que mide la ganancia de precisión, y tal y como se puede observar, es posible que la aticipidad de una regla sea inferior a cero cuando la regla tiene una baja calidad ya que el porcentaje de ejemplos negativos cubiertos es superior al de los positivos. En este sentido, una regla de interés para el experto debería tener siempre un valor positivo, es decir, se debe cumplir la siguiente desigualdad para obtener una regla de interés:

$$\frac{p}{p+n} > \frac{P}{P+N} \quad (4)$$

Si descomponemos esta desigualdad [1] podemos decir que para que una regla obtenga una atipicidad positiva y ser un subgrupo de interés, obtenemos que el producto escalar entre $p\overline{n}$ debe ser superior que $\overline{p}n$. En consecuencia, esta descomposición nos indica que el producto de ejemplos cubiertos y no cubiertos correctamente debería ser superior al producto de ejemplos cubiertos y no cubiertos incorrectamente:

$$p\ \overline{n} > \overline{p}\ n \quad (5)$$

En el trabajo publicado en [1] se presenta tras este primer análisis, la compatibilidad entre el descubrimiento de subgrupos, los conjuntos de contraste y los patrones emergentes gracias al uso de esta medida de calidad.

Por ejemplo, en el caso de los patrones emergentes (según definición) podemos afirmar que una regla es emergente cuando su índice de crecimiento es superior a 1, es decir:

$$GR(R) = \frac{\frac{p}{P}}{\frac{n}{N}} > 1 \quad (6)$$

Si aplicamos una descomposición a esta formulación descubrimos que:

$$\frac{\frac{p}{P}}{\frac{n}{N}} > 1$$
$$p\ \overline{n} > \overline{p}\ n \quad (7)$$

Según esta demostración podemos afirmar que la descomposición de la atipicidad positiva y del índice de crecimiento son similares, es decir, si un subgrupo es de interés también es un patrón emergente. Además, por otro lado se puede observar una relación directa entre la atipicidad y la diferencia de soporte en [2] que dice:

$$DS(R) = \frac{WRAcc(R)}{p(PIS) \cdot p(NIS)} \geq \delta \quad (8)$$

donde $p(PIS)$ es el porcentaje de ejemplos positivos para la clase del conjunto de datos a analizar y $p(NIS)$ es el porcentaje de ejemplos negativos para el conjunto de datos.

Al mismo tiempo, es importante tener en cuenta que el dominio de la atipicidad para un problema depende del porcentaje de los ejemplos para la clase o variable objetivo [7], y por tanto, el dominio se puede determinar mediante las siguientes ecuaciones para el límite inferior (LB_{WRAcc}) y para el superior (UB_{WRAcc}):

$$LB_{WRAcc} = (1 - p(PIS)) \cdot (0 - p(PIS)) \quad (9)$$

$$UB_{WRAcc} = p(PIS) \cdot (1 - p(PIS)) = p(PIS) \cdot p(NIS) \quad (10)$$

Es decir, el valor de DS depende de las propiedades del problema ya que está directamente relacionado con el porcentaje de ejemplos de la clase a analizar, y en resumen:

$$DS(R) = \frac{WRAcc(R)}{p(PIS) \cdot p(NIS)} = \frac{WRAcc(R)}{UB_{WRAcc}} \geq \delta \quad (11)$$

Una regla se considera conjunto de contraste con un $\delta = 0{,}10$ cuando:

$$WRAcc(R) \geq 0{,}10 \cdot UB_{WRAcc} \quad (12)$$

Sin embargo, tenemos la necesidad de estandarizar este valor de $WRAcc$ a $WRAcc$ normalizada ($WRAccN$) ya que debemos evitar la utilización de un valor δ que esté condicionado a este porcentaje de ejemplos de la clase a analizar. Esta mejora se consigue con respecto a la homogeneización de esta medida clave dentro del descubrimiento de reglas descriptivas supervisadas. La $WRAccN$ se normaliza en el intervalo [0,1] mediante la siguiente ecuación donde una regla es subgrupo de interés cuando su valor sea superior a 0.5.

$$WRAccN(R) = \frac{WRAcc(R) - LB_{WRAcc}}{UB_{WRAcc} - LB_{WRAcc}} \quad (13)$$

Considerando un δ igual a 0.10, y el intervalo positivo de $WRAccN$ igual a (0.5, 1.0], una regla se considera conjunto de contraste con un $\delta = 0{,}10$ cuando:

$$WRAccN(R) \geq 0{,}50 + 0{,}10 \cdot (1{,}00 - 0{,}50)$$
$$WRAccN(R) \geq 0{,}55 \quad (14)$$

, es decir, cuando el valor de la $WRAccN$ es superior o igual que el 10 % de la parte positiva de $WRAcc$.

En conclusión, la atipicidad es un factor clave a medir dentro del descubrimiento de reglas descriptivas supervisadas donde una regla con un valor de $WRAccN$ superior a 0.50 se considera de interés para el descubrimiento de subgrupos, y emergente para la extracción de patrones emergentes. Además, si la $WRAccN$ es superior o igual a 0.55 se considera también regla de contraste. Esta demostración muestra la relación directa entre todas estas técnicas que se han agrupado dentro del descubrimiento de reglas descriptivas supervisadas y muestra la importancia de medir la calidad de una regla con respecto a la $WRAcc$ en un problema que se desea analizar. Por otro lado, es interesante indicar que solo con el análisis de esta medida de calidad, los expertos serán capaces de determinar si la regla representa un subgrupo, una regla emergente y/o una regla de contraste. De la misma forma, esta medida debe ser clave para el diseño de nuevas propuestas dentro del descubrimiento de reglas descriptivas supervisadas.

Agradecimientos

Este trabajo ha sido subvencionado por el Ministerio de Economía y Competitividad de España bajo el proyecto TIN2014-916 57251-P y TIN2015-68454-R (Fondos FEDER).

Referencias

[1] C. J. Carmona, M. J. del Jesus, and F. Herrera, "A Unifying Analysis for the Supervised Descriptive Rule Discovery via the Weighted Relative Accuracy," *Knowledge-Based Systems*, vol. 139, pp. 89–100, 2018.

[2] P. Kralj-Novak, N. Lavrac, and G. I. Webb, "Supervised Descriptive Rule Discovery: A Unifying Survey of Constrast Set, Emerging Pateern and Subgroup Mining," *Journal of Machine Learning Research*, vol. 10, pp. 377–403, 2009.

[3] W. Kloesgen, "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996, pp. 249–271.

[4] G. Z. Dong and J. Y. Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," in *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 1999, pp. 43–52.

[5] S. D. Bay and M. J. Pazzani, "Detecting group differences: Mining contrast sets," *Data Mining and Knowledge Discovery*, vol. 5, no. 3, pp. 213–246, 2001.

[6] N. Lavrac, P. A. Flach, and B. Zupan, "Rule Evaluation Measures: A Unifying View," in *Proc. of the 9th International Workshop on Inductive Logic Programming*, ser. LNCS, vol. 1634. Springer, 1999, pp. 174–185.

[7] C. J. Carmona, V. Ruiz-Rodado, M. J. del Jesus, A. Weber, M. Grootveld, P. González, and D. Elizondo, "A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans," *Information Sciences*, vol. 298, pp. 180–197, 2015.