Análisis Preliminar del Rendimiento de Algoritmos para el Procesos de Descubrimiento de Reglas de Pertenencia a Grupos

Gabriel Ciciliani [1], Sebastián Martins [2], Hernán Merlino [1,2]

[1] Programa de Magíster en Ingeniería en Sistemas de información. Escuela de Posgrado. Facultad Regional Buenos aires. Universidad Tecnológica Nacional, Argentina
[2] Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús. Argentina
gabriel.ciciliani@gmail.com, smartins089@gmail.com, hmerlino@gmail.com

Resumen. En el campo de la explotación de información, el proceso de descubrimiento de reglas de pertenencia a grupos se caracteriza por la utilización combinada de un proceso de descubrimiento de grupos (clustering) y uno de inducción de reglas. Dada la variedad de algoritmos de clustering e inducción de reglas disponibles en la actualidad, es de interés poder conocer a priori qué pareja de algoritmos es más conveniente para un set de datos, en a base sus características. En este artículo se propone un diseño experimental que permita validar el rendimiento de los algoritmos, en base a métricas internas, para distintos tipos de sets de datos, con características específicas, de forma tal que permita vincular dichas características con la pareja de algoritmos que mejor rendimiento ofrece. En adición, se presentan resultados preliminares obtenidos.

Palabras Clave. Minería de datos, clustering, inducción de reglas, descubrimiento de reglas de pertenencia a grupos, estudio de algoritmos

1 Introducción

La **Explotación de Información** es una sub-disciplina de los Sistemas de Información, que brinda a la Inteligencia de Negocio las herramientas para la transformación de información en conocimiento (García-Martínez et al., 2015) y se define como la búsqueda de patrones interesantes y de reglas importantes, previamente desconocidas, en grandes cantidades de información almacenada en distintos medios (Martins, 2016). Un **proceso de Explotación de Información** consiste en un grupo de tareas relacionadas que se realizan con el objetivo de obtener información útil y significativa a partir de grandes cantidades de datos. Para esto, dichos procesos se valen de la utilización de algoritmos de Minería de Datos (García-Martínez et al., 2015; Martins, 2016).

Un procedimiento recurrente en la explotación de información es el **proceso de descubrimiento de reglas de pertenencia a grupos** (García-Martínez et al., 2015) consiste en tomar el conjunto de datos a estudiar y aplicar un algoritmo de agrupamiento o "clustering" para separarlo en distintos grupos (también llamados clases o clusters) y aplicar luego algoritmos de inducción de reglas para explicitar las

combinaciones de atributos, con sus respectivos rangos de valores, que definen la pertenencia a cada grupo descubierto (Kaski, 1997; Hall y Holmes, 2003).
Tanto para el procedimiento de *clustering* como para la de inducción de reglas existen varios algoritmos (Xu y Tian 2015; Sehgal y Garg 2014; Panchuk, 2015). Esto se explica por la necesidad de encontrar, en la diversidad de dominios de negocio, los algoritmos de explotación de información que mejor identifican los patrones ocultos en dicha masa de información.
Este trabajo se estructura en 5 secciones: la presente sección (1) ofrece una introducción, la sección (2) presenta una descripción de la problemática a abordar, la sección (3) consiste en la formulación del diseño experimental, la sección (4) muestra los resultados preliminares, finalizando con las conclusiones en la sección (5).

2 Descripción del Problema

Numerosos trabajos demuestran que la performance de los algoritmos varía notablemente tanto con las características del set de datos a analizar como por los valores de los ejemplos de dicho set (Kogan, 2007; López-Nocera, 2012; Panchuk, 2015; Sehgal y Garg 2014; Smith, Woo, Ciesielski y Ibrahim 2002). Podemos decir entonces que uno de los desafíos de la ingeniería de explotación de información es encontrar los algoritmos que mejor describen el set de datos a analizar.

Varios investigadores se han volcado al estudio de la relación entre las características de un set de datos y la performance de algoritmos relevantes en el descubrimiento de grupos (Xu y Tian 2015), inducción de reglas (Smith et al., 2002) y el descubrimiento de reglas de pertenencia a grupos (Kogan, 2007; López-Nocera, 2012; Panchuk, 2015)

Particularmente para este último proceso, los trabajos citados proponen una clasificación de los dominios a los cuales pertenecen los datos analizados y demuestran empíricamente que pareja de algoritmos es más eficaz para cada tipo de dominio. Dicha clasificación se basa en parámetros del mismo y del set de datos a analizar, como la cantidad de atributos, la cantidad de clases (grupos), la cantidad de reglas por clase, etc. En este contexto, puede apreciarse que parte de las variables utilizadas para el estudio están estrechamente vinculadas con los resultados del proceso de descubrimiento de reglas de pertenencia a grupos en sí.

Partiendo de los estudios realizados en (Kogan, 2007; López-Nocera, 2012; Panchuk, 2015) y en base a las líneas de trabajo futuras definidas, se desprende la necesidad de extender el estudio en busca de un método que permita identificar a priori qué combinación de algoritmos de clustering e inducción de reglas es el más adecuado para un set de datos dado, sin recurrir a características que dependen en sí mismas de los resultados del proceso de descubrimiento de reglas de pertenencia a grupos.

En síntesis, el presente trabajo pretende sentar las bases del estudio que permita predecir la mejor conformación de un proceso de descubrimiento de reglas de pertenencia a grupos únicamente en base a características específicas del set de datos a analizar.

3 Formulación del Diseño Experimental

Para abordar el problema definido en la sección anterior, se realizará un diseño experimental que permita comprender el comportamiento de los algoritmos de acuerdo a las características del set de datos, generando una base de conocimiento que permita posteriormente definir las parejas más eficientes.
Esto implica, en primera instancia, definir la forma mediante la cual se determinará la eficiencia del proceso. En (Feldman y Sanger, 2007) se identifican tres alternativas posibles:

- *Evaluación externa*: consiste en partir de un set de datos conocido y, aplicando diferentes parejas de algoritmos, encontrar cual es el que mejor describe la realidad. La limitación de esta estrategia es evidenciada por el objetivo del proceso de descubrimiento de grupos en sí: la búsqueda de patrones ocultos en el set de datos. No siempre se cuenta con una verdad conocida en base a la cual se pueden evaluar los distintos algoritmos.
- *Evaluación manual*: un humano experto analiza los resultados generados por cada pareja de algoritmos. La debilidad de esta estrategia radica en los tiempos/costes y la subjetividad de dichas revisiones manuales, especialmente si se considera que en muchos casos los resultados obtenidos son contra-intuitivos.
- *Evaluación Interna*: se caracteriza por definir métricas o índices que de alguna forma describan la *calidad* de los resultados obtenidos. Esta estrategia cuenta también con cierto grado de subjetividad: buenos valores de métricas no siempre significan alto valor de calidad en los resultados obtenidos. Además, existe una tendencia de ciertos algoritmos a generar buenos valores de métricas relacionadas (Van Craenendonck y Blockeel, 2015).

A pesar de dichas limitaciones, la evaluación interna no requiere conocimientos previos de los patrones que se intentan extraer de los datos, y el cálculo y evaluación de métricas puede automatizarse. De esta forma, se puede aplicar a cualquier problema de descubrimiento de reglas de pertenencia a grupos. Para mitigar la tendencia de ciertos algoritmos de *clustering* a puntuar mejor en ciertas métricas, se utilizaron cinco índices de diferente naturaleza, detallados en la sección 3.3. Como criterios adicionales, se incluyeron también métricas secundarias que permiten una segunda instancia de evaluación en caso de empate. Para la evaluación de los algoritmos de inducción de reglas, se aplica una estrategia similar, aunque utilizando una única métrica primaria y una única secundaria, descritas en la sección 3.5.
Como se describió en la sección anterior, el objetivo del experimento es poder asociar características de un set de datos con una pareja de algoritmos que retorne la mejor separación en grupos y las mejores reglas que describen esos grupos, en base a métricas internas. Para ello se siguen los siguientes pasos:

1) Generar un set de datos artificialmente con valores específicos para cada una de las características a estudiar
2) Verificar que se cumplan los valores establecidos para cada característica

3) Someterlo a un proceso de descubrimiento de reglas de pertenencia a grupos utilizando diferentes parejas de algoritmos
4) Calcular métricas internas tanto para la fase de separación en grupos como para la de inducción de reglas
5) Determinar la pareja ganadora en base a las métricas obtenidas
6) Analizar patrones en el comportamiento de las parejas de algoritmos asociados con las características de los set de datos generadas en la masa de datos experimentales.

La secuencia anterior se repite para cada set de datos, tipificado de la A a la Y (tabla 1) de acuerdo con la combinación de características (descriptas en la sección 3.1). Una vez obtenidas las relaciones entre las características del set de datos y los algoritmos que mejores resultados generan, podría ser posible determinar qué pareja de algoritmos utilizar con solo conocer ciertas características del set, sin necesidad de incurrir en un análisis comparativo de algoritmos. A continuación se describe cada fase del experimento en detalle.

3.1 Generación y validación del set de datos

Los sets de datos para este estudio son generados artificialmente y manipulados de forma que posean las características requeridas. La estructura es del tipo matricial, donde cada ejemplo del set está constituido por una cantidad fija de valores numéricos reales. La figura 1 presenta la estructura del subproceso de generación del set de datos.

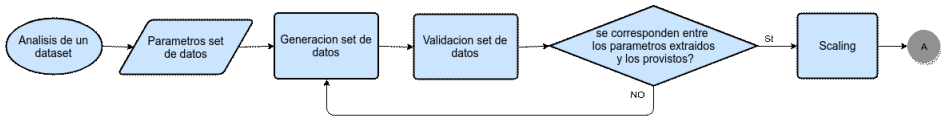

Fig 1. Diagrama de flujo subproceso de generación del set de datos

En cuanto a la distribución de los diferentes atributos, se optó por intercalar atributos "normales" con atributos "uniformes" considerando que, en la práctica, los datos no son aleatorios y responden generalmente a distribuciones conocidas. Las características del set de datos a considerar en este estudio son:

a) **Cantidad de atributos (CA):** cantidad de columnas o valores que presenta cada ejemplo del set de datos.
b) **Cantidad de ejemplos (CE):** número de filas que presenta el set de datos.
c) **Porcentaje de ejemplos con tendencia lineal (PL):** para determinar si a un ejemplo se lo considera dentro de un grupo lineal se calculará la distancia del mismo a la recta de ajuste obtenida por el método de descomposición en valores singulares (Golub, Reinsch 1970). La distancia umbral está definida como un porcentaje del módulo de la recta contenida entre los planos correspondientes a los máximos valores de cada coordenada.
d) Porcentaje de ejemplos repetidos (PR): cantidad de ejemplos que tienen exactamente los mismos valores en todos sus atributos.
e) Cantidad de grupos de ejemplos repetidos (GR): ligado directamente al punto anterior, esta característica indica cuántos grupos de ejemplos repetidos existen.

f) Porcentaje de *outliers* (PO): se consideran atípicos o *outliers* a aquellos ejemplos dentro del 20% más lejano al punto medio de la nube de puntos cuya distancia sea mayor a la distancia del ejemplo inmediatamente anterior a dicho 20%, multiplicado por un coeficiente.

Variando la composición de cada set en base a las características mencionadas se obtienen los diferentes tipos de sets a estudiar (A-Y) . La tabla 1 ilustra las combinaciones realizadas.

Tabla 1. Combinaciones de atributos utilizadas para la generación de datos.

Tipo	CA	CE	PL	PR	GR	PO
A	8	1000	N	N	0	N
B	8	1000	B	N	0	N
C	16	1000	B	N	0	N
D	24	1000	B	N	0	N
E	8	1000	M	N	0	N
F	8	1000	A	N	0	N
G	8	1000	N	B	2	N
H	8	1000	N	M	2	N
I	8	1000	N	A	2	N
J	8	1000	N	M	3	N
K	8	1000	N	A	3	N
L	8	1000	N	M	4	N
M	8	1000	N	A	4	N

Tipo	CA	CE	PL	PR	GR	PO
N	8	1000	N	N	0	B
O	8	1000	N	N	0	M
P	8	1000	N	N	0	A
Q	8	1000	B	B	2	B
R	8	1000	B	B	2	M
S	8	1000	B	B	2	A
T	8	1000	B	A	2	B
U	8	1000	B	A	2	M
V	8	1000	B	A	2	A
W	8	1000	A	B	2	B
X	8	1000	A	B	2	M
Y	8	1000	A	B	2	A

A = Alto, M = Medio, B = Bajo, N = nulo

Característica	B	M	A
PL	10%	40%	80%
PR	10%	20%	40%
PO	6%	12%	18%

Antes de someter al set de datos generado a los diferentes algoritmos, es necesario validar que las características solicitadas estén presentes utilizando el mismo método con el que luego se pretende caracterizar a los sets reales.
Si una o más características, deducidas por la rutina de análisis, difieren de las especificadas al generar el set de datos, se genera un set nuevo hasta que no haya discrepancias o se exceda una cantidad límite de intentos. Cabe aclarar que la fase de análisis contempla un margen de error de +/- 5% para las métricas expresadas como porcentajes de la cantidad de ejemplos (PL, PR y PO).
Este proceso de validación es el mismo que se utilizará posteriormente para predecir qué pareja de algoritmos es la más conveniente en base a las características del set de datos obtenidas.

3.2 Descubrimiento de grupos

Una vez generado y validado el set de datos, se lo somete a diferentes algoritmos de descubrimientos de grupo o *clustering*. La figura 2 ilustra el procedimiento aplicado.

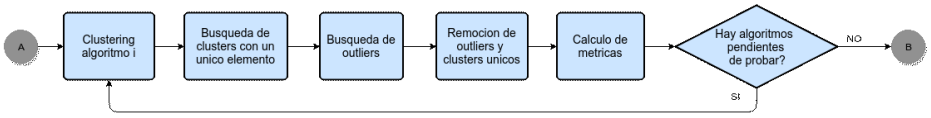

Fig 2. Diagrama de flujo subproceso de descubrimiento de grupos

Los algoritmos y variantes utilizados en este trabajo son:

1) K-Means en tres variantes de inicialización de centroides distintas:
   a) Definición de centros aleatoria,
   b) Variante k-means++ (Arthur, Vassilvitskii 2007)
   c) Determinación de centros mediante Análisis de componentes principales (Alrabea, Senthilkumar, Al-Shalabi, Bader 2013);
2) DBSCAN
3) Birch
4) Meanshift.

Para los algoritmos que no poseen una estrategia automática de configuración de sus parámetros, se utiliza un estrategia del tipo *grid-search* para determinar los parámetros de operación óptimos del estimador.

Dado que no todos los algoritmos operan de la misma forma, pueden darse situaciones donde el algoritmo *ignore* ciertos ejemplos por no poder asociarlos a un *cluster* o donde algunos de los *clusters* generados posea un único elemento. Estas dos situaciones se consideran "indeseadas" ya que se espera que ningún dato sea omitido en el análisis y un cluster de un único elemento no constituye un grupo en sí. Es por esto que tanto los grupos únicos como los ejemplos no clasificados son ignorados tanto en el cálculo de métricas como en la etapa de inducción de reglas, previo almacenamiento de la cantidad de ejemplos omitidos y el motivo, para ser considerado en la selección del algoritmo ganador, de ser necesario.

3.3 Cálculo de métricas sobre los grupos obtenidos

Los grupos descubiertos por cada algoritmo y variante mencionados son puntuados utilizando las métricas siguientes **métricas primarias**: a) Índice Davies-Bouldin, b) Índice Dunn, c) Índice Calinski-Harabasz, d) Índice silhouette y e) Suma de cuadrados. Como **métricas secundarias** se utilizan: tiempo de cómputo, cantidad de clusters de un único elemento, y puntos descartados por el algoritmo. La figura 3 resume el proceso de determinación del algoritmo ganador.

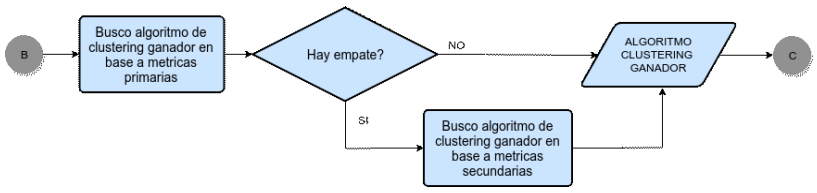

Fig 3. Diagrama de flujo subproceso determinación de algoritmo ganador (clustering)

Para obtener el algoritmo de descubrimiento de grupos ganador, se busca aquel que mejores valores tuvo para la mayor cantidad de métricas primarias. El valor de la métrica de ser al menos un 5% mayor (o menor dependiendo de la métrica) a la actual ganadora para tomar su lugar. Si existe un empate entre dos o más algoritmos, se utilizan las métricas secundarias.
Las métricas de ejemplos ignorados y grupos de un único elemento se utilizan solo si algún algoritmo finalista incurrió en dichos comportamientos. En caso de que exista un empate nuevamente, se elige uno de los algoritmos ganadores al azar para proceder a la etapa de inducción de reglas, dejando registro de los finalistas.

3.4 Inducción de reglas de pertenencia a grupos

En esta etapa se utilizan las etiquetas generadas por el algoritmo ganador para inducir reglas que describen las características de cada grupo. La figura 4 resume el subproceso de inducción de reglas de pertenencia a grupos. Los algoritmos de inducción de reglas utilizados para este trabajo son: CART y CN2. Para la determinación del valor mínimo de ejemplos por hoja óptimo, se utiliza también una estrategia del tipo Grid-search, combinado con validación cruzada o *cross-validation*, para evitar sesgos producidos por la utilización del mismo set de datos tanto en el entrenamiento como para en la evaluación.

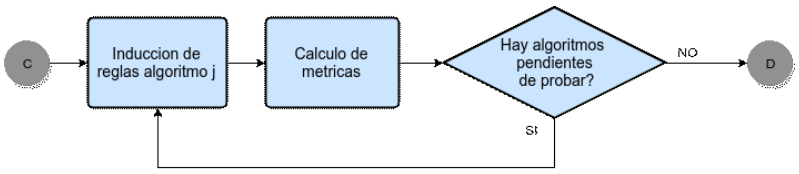

Fig 4. Diagrama de flujo subproceso inducción de reglas

3.5 Cálculo de métricas sobre las reglas generadas

Para determinar el algoritmo de inducción de reglas ganador se utiliza la curva ROC (*Receiver Operating Characteristic* o Característica Operativa del Receptor). Esta, a diferencia de otras métricas generalmente utilizadas, permite contemplar los falsos positivos, factor vital en contextos de desbalanceo de datos. Investigaciones recientes (Jurgovsky 2018) presentan estrategias para evaluar la performance de un clasificador basadas en el área bajo la curva (AUC, por sus siglas en ingleś) ROC. En este trabajo, una vez calculados los puntos de la curva ROC, se computa el área bajo la misma utilizando la regla del trapecio. A mayor valor de área, mejor será la capacidad de las reglas generadas de clasificar los ejemplos correctamente (Bradley, A 1997). La figura 5 ilustra el flujo del subproceso determinación de algoritmo de inducción de reglas ganador.
El área bajo la curva ROC es, entonces, la única métrica primaria utilizada para la evaluación de los algoritmos de inducción. El tiempo de cómputo será utilizado como métrica secundaria en caso de empate.

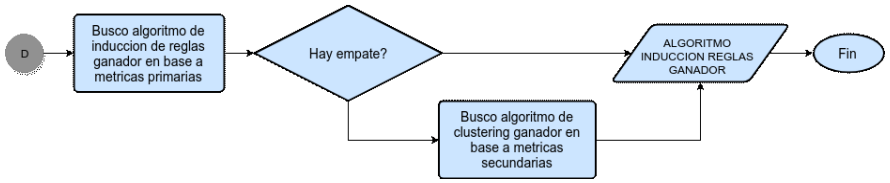

Fig 5. Diagrama de flujo subproceso determinación de algoritmo ganador (inducción de reglas)

4. Resultados preliminares

En la tabla 2 podemos observar los resultados obtenidos en base al primer lote de ejecuciones (417) del proceso realizadas hasta el momento (de un total de 1300), para cada uno de los 25 tipos de set de datos. Las filas resaltadas en gris oscuro corresponden a aquellos tipos que respondieron mejor al algoritmo ganador en más de un 70% de los casos.

En general puede observarse una baja cantidad de algoritmos ganadores determinados por tiempo de cómputo, aunque para el tipo A (ver tabla 1), ocurre aproximadamente en el 22% de los casos, siendo éste el porcentaje más alto del lote de datos. Asimismo, para los tipos resaltados, se observa que en su mayoría los algoritmos se decidieron en base a la cantidad de métricas sobresalientes. Puede observarse también un porcentaje de empates de un dígito o menos para todos los sets de datos. Ni los clusters de un único elemento, ni los ejemplos ignorados fueron un factor relevante para elegir el algoritmo ganador, como puede inferirse de los valores de la columna %CUE y %IE.

Los resultados de las corridas para los algoritmos de inducción arrojó que en el 100% de los casos, el algoritmo ganador fue CART y se destacó por tiempo de cómputo y no por presentar un área bajo la curva ROC mayor a CN2. Solo en el 2.8% de los casos, contemplando todos los tipos de sets de datos, CN2 fue elegido como ganador, en todos los casos por mostrar un mejor valor de área.

5 Conclusiones

De los resultados preliminares puede concluirse que para los tipos de datos resaltados en la tabla 2, existe una fuerte tendencia de los algoritmos ganadores a realizar una separación de los ejemplos en grupos de mejor *calidad* desde la perspectiva de las métricas internas utilizadas. Esto sugiere que el utilizar dicho algoritmo en un set de datos, con características similares a las tipificadas, tiene altas probabilidades de generar mejores grupos, comparado con los generados por los algoritmos restantes.

En cuanto al algoritmo utilizado para inducir las reglas de pertenencia a grupos, CART supera a CN2 en más del 97 de los casos, pero únicamente en performance tornandolo el algoritmo a elegir, sin importar el tipo de datos o el algoritmo de clustering utilizado en la etapa anterior.

Como futuro trabajo se prevé ampliar el análisis sobre un conjunto de datos mayor en búsqueda de patrones adicionales que vinculen características del set de datos, con las métricas y los algoritmos. Asimismo se pretende validar empíricamente los patrones obtenidos utilizando sets de datos reales. Se proponen como futuras líneas de

investigación la inclusión de nuevas características del set de datos a considerar y la ampliación de algoritmos tanto de clustering como de inducción de reglas.

Agradecimientos

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por los Proyectos 80020160400001LA y 80020160500002LA de la Secretaría de Ciencia y Tecnología de la Universidad Nacional de Lanús.

Tabla 2. Resultados de las ejecuciones para los algoritmos de descubrimiento de grupos.

Tipo	1er Alg	% casos 1	2do Alg	% casos 2	% mp	% tiempo	% CUE	% EI	% emp
A	kmeans_++	78,5	kmeans_random	10,8	77,8	22,2	0	0	0
B	kmeans_++	69,8	kmeans_random	16,3	80,4	19,6	0	0	0
C	kmeans_++	48,6	kmeans_random	34,9	98,1	1,9	0	0	0
D	kmeans_++	44,6	kmeans_random	33	97,9	2,1	0	0	0
E	meanshift	58,7	birch	28,3	100	0	0	0	0
F	dbscan	59,9	kmeans_++	18,4	100	0	0	0	1,2
G	kmeans_++	70,9	kmeans_random	13	83,3	16,7	0	0	0,5
H	kmeans_++	53,6	dbscan	31,8	88,5	11,5	0	0	0,5
I	dbscan	68,4	kmeans_++	25	95,9	4,1	0	0	0
J	kmeans_++	60,7	dbscan	20,1	89,2	10,8	0	0	2,8
K	dbscan	87,9	kmeans_++	9,7	100	0	0	0	0,5
L	kmeans_++	65	dbscan	13,7	89,8	10,2	0	0	3,8
M	dbscan	90,4	kmeans_++	7,6	100	0	0	0	3,8
N	meanshift	92,2	kmeans_++	3,8	100	0	0	0	0
O	meanshift	92	kmeans_++	3,3	100	0	0	0	0
P	meanshift	86,1	kmeans_++	5,7	100	0	0	0	0
Q	meanshift	90,3	birch	5,7	100	0	0	0	0
R	meanshift	86,7	kmeans_++	5,5	100	0	0	0	0
S	meanshift	82,3	kmeans_++	8,6	100	0	0	0	0
T	dbscan	100	N/A	0	100	0	0	0	0
U	dbscan	99,8	meanshift	0,2	100	0	0	0	0
V	dbscan	99,5	meanshift	0,5	100	0	0	0	0
W	meanshift	53,3	birch	21,8	100	0	0	0	0
X	meanshift	56,5	birch	22	100	0	0	0	0
Y	meanshift	57,1	kmeans_++	20,6	100	0	0	0	0

***Referencias**: **1er Alg**: Algoritmo ganador; **% casos 1**: Porcentaje del set de datos que respondió mejor al algoritmo ganador; **2do Alg**: Segundo algoritmo en cantidad de set de datos mejor agrupados; **% casos 2**: Porcentaje del set de datos correspondiente al segundo algoritmo; **% mp**: Porcentaje de casos en los que el 1er algoritmo ganó en base a las métricas primarias; **% tiempo**: Porcentaje de casos en los que el 1er algoritmo ganó por tiempo de cómputo; **% CUE**: Porcentaje de casos en los que el 1er algoritmo ganó por haber generado menor cantidad de clusters de un único elemento; **% EI**: Porcentaje de casos en los que el 1er algoritmo ganó por haber ignorado una menor cantidad de ejemplos; **% emp**: Porcentaje del total de sets de datos analizado donde existió un empate entre dos o más algoritmos.*

Referencias

Alrabea, A., Senthilkumar, A. V., Al-Shalabi, H., & Bader, A. (2013). Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with PCA. Journal of Advances in Computer Networks, 1(2), 137-142.

Arthur, D., Vassilvitskii S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

Cogliati, M., Britos, P. y García Martínez, R. (2006). Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874. http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf

Feldman, Ronen; Sanger, James (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge Univ. Press. ISBN 0521836573. OCLC 915286380

García-Martínez, R., Britos, P., Martins, S., & Baldizzoni, E. (2015). Explotación de Información. Ingeniería de Proyectos. Editorial Nueva Librería ISBN, 978-987-1871-34-6.

Golub, G. H.; Reinsch, C. (1970). "Singular value decomposition and least squares solutions". Numerische Mathematik. 14 (5): 403–420. doi:10.1007/BF02163027. MR 1553974.

Grosser, H., Britos, P. y García Martínez, R. (2005). Detecting Fraud in Mobile Telephony Using Neural Networks

Hall, M. y Holmes, G. (2003) Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, Tomo 6, páginas 1437-1447.

Jurgovsky, Johannes, et al. "Sequence classification for credit-card fraud detection." Expert Systems with Applications 100 (2018): 234-245.

Kaski, S. (1997). Data exploration using self-organizing maps.

Kogan A., (2007). Integración de Algoritmos de Inducción y Agrupamiento. Estudio del Comportamiento.

López-Nocera M. (2012). Descubrimiento De Conocimiento Mediante La Integración De Algoritmos De Explotación De La Información.

Martins, S., Pesado, P., & García-Martínez, R. (2016). Intelligent Systems in Modeling Phase of Information Mining Development Process. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 3-15). Springer International Publishing.

Panchuk J. (2015) Comportamiento De Integración De Algoritmos Para Descubrimiento De Reglas De Pertenencia A Grupos. http://sistemas.unla.edu.ar/sistemas/gisi/TFLS/Panchuk-TFL.pdf

Sehgal, G., & Garg, D. K. (2014). Comparison of Various Clustering Algorithms. International Journal of Computer Science and Information Technologies, 5(3), 3074-307.

Smith, K. A., Woo, F., Ciesielski, V., & Ibrahim, R. (2002). Matching data mining algorithm suitability to data characteristics using a self-organizing map. In Hybrid information systems (pp. 169-179).

Van Craenendonck, T., & Blockeel, H. (2015). Using internal validity measures to compare clustering algorithms. In Benelearn 2015 Poster presentations (online) (pp. 1-8).

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.