# Métodos de agrupamiento LA & SIA: Comparación computacional

Mauricio Naranjo[a], Rubén Pazmiño[b], Miguel Conde[c] , Francisco Peñalvo[d]
[a] Pontificia Universidad Católica del Ecuador sede Ambato, Programa Magister en Ciencias de la Educación.
[b] Facultad de Ciencias, CIDED, Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador
[c] Department of Computer Science, University of León, León, Spain
[d] Department of Computer Science, University of Salamanca, Salamanca, Spain
maurysnaranjo@gmail.com, rpazmino@espoch.edu.ec, miguel.conde@unileon.es, fgarcia@usal.es

***Resumen*—Las analíticas de aprendizaje son y siguen siendo una tecnología emergente según el informe Horizon del 2016, es por ello por lo que el estudio y la búsqueda de nuevas técnicas de análisis es importante. El análisis Estadístico Implicativo permite descubrir R-reglas de la forma a → b sobre un conjunto de variables o sujetos. Las reglas se representan gráficamente por dendogramas que sirven como una nueva herramienta clustering basada en el concepto de cohesión. El Clustering es una de las técnicas más utilizadas en Learning Analytics para la exploración de datos para condensarlos en grupos heterogéneos de objetos similares entre sí. El objetivo de este artículo es comparar desde la ocupación de memoria las funciones cluster utilizadas en Learning Analytics hclust.vector, dendro.variables y diana y las funciones callHierarchyTree y callSimilarityTree utilizadas en el Análisis Estadístico Implicativo, esto permitirá determinar las ventajas de las nuevas técnicas de análisis cluster basadas en el Análisis Estadístico Implicativo. El análisis comparativo se realizó mediante un diseño cuasi experimental bifactorial del tipo RGXO, para controlar las variables exógenas se utilizó similares arquitecturas de hardware (Procesador Core I7, Velocidad 2,2 Ghz y Memoria RAM 8Gb) y de software (Windows 8, Ubuntu 16.04, MacOS Sierra 10.12, R v3.4.1 y RStudio v1.0.153). El colectivo de estudio estuvo conformado por 100000 bases de datos de 1000 observaciones y 100 variables dicotómicas. Se utilizó un método de muestreo aleatorio simple, el tamaño de muestra utilizado fue de 383 bases de datos. Los resultados demuestran que no existe diferencia significativa en la ocupación de memoria entre los métodos simlrty, dendro_diana y hclust_vector, es decir estadísticamente los tres métodos son equivalentes y son los que menos memoria ocupan. El método hrarchy ocupa el segundo lugar en mayor ocupación de memoria y el método que más memoria utiliza es dendro_variables.***

***Palabras Claves*—analíticas de aprendizaje, análisis estadístico implicativo, métodos cluster, callSimilarityTree, dendro.variables, RCHIC**

***Abstract*—Learning analytics are and continue being an emerging technology according to the 2016 Horizon report, that is why the study and the new search for analysis techniques is important. Statistical Implicative Analysis allows to discover R-rules of the form a → b on a set of variables or subjects. The rules are represented graphically by dendograms that serve as a new clustering tool based on the concept of cohesion. Clustering is one of the most used techniques in Learning Analytics to explore data for condense them into heterogeneous groups for similar objects for each other. The objective of this article is to compare the memory functions of the cluster functions used in Learning Analytics hclust.vector, dendro.variables and diana and the functions callHierarchyTree and callSimilarityTree used in Statistical Analysis Implicative, this will allow to determine the advantages of the new techniques of cluster analysis based on Statistical Implicative Analysis. The comparative analysis was carried out through a quasi-experimental bifactorial design of the RGXO type, to control the exogenous variables, similar hardware architectures were used (Core I7 processor, 2.2 Ghz velocity and 8Gb RAM memory) and software (Windows 8, Ubuntu 16.04, MacOS Sierra 10.12, R v3.4.1 and RStudio v1.0.153). The population consisted of 100000 databases of 1000 observations and 100 dichotomous variables. A simple random sampling method was used, the sample size used was 383 databases. The results show that there is no significant difference in memory occupancy between the methods simlrty, dendro_diana and hclust_vector, that is, statistically all three methods are equivalent and are the ones with the least memory. The hrarchy method occupies the second highest memory allocation and the method that uses the most memory is dendro_variables.***

***Keywords*—learning analytics, statistical implicative analysis, cluster methods, callSimilarityTree, dendro.variables, RCHIC**

## I. Introducción

El análisis de datos es fundamental en los procesos educativos de deserción, rendimiento, seguimiento, rendimiento, uso de tecnologías, entre otros; al no utilizar las técnicas más adecuadas se obtiene obstrucción, receso, estancamiento y lentitud en los cálculos lo que hace que las técnicas de análisis de datos sean inaplicables. Es necesario utilizar técnicas óptimas que minimicen el espacio de memoria y el tiempo de procesamiento (complejidad algorítmica). Por esta razón esta investigación determina las técnicas óptimas desde el punto de vista de la ocupación de memoria utilizadas en el Análisis Estadístico Implicativo (SIA) y Learning Analytics (LA).

El Análisis Estadístico Implicativo se desarrolló al encontrar problemas o cuestiones planteadas [1], su objetivo es estructurar los datos, a través de técnicas comunes de adquisición de conocimientos en cualquier proceso de aprendizaje [2]. Por otra parte, el análisis estadístico Implicativo según la investigación realizada por [3], indica que su primera aplicación es el ámbito educativo, específicamente en el área de la matemática. Del estudio realizado se desprende que los artículos sobre Educación (71 artículos) duplican aquellos de desarrollo teórico (27 artículos), por lo cual el investigador concluye que existe muchas experiencias en la aplicación en el área educativa, con lo cual el autor motiva a los educadores a utilizar esta nueva técnica estadística multivariada. Además, nos hace ver la compatibilidad que existe entre las técnicas de análisis de datos del SIA y del LA [4]. El Análisis Estadístico Implicativo se automatiza mediante el software CHIC en su versión propietario desarrollado por el profesor Raphael

Couturier [5] y en su versión libre llamado RCHIC [6]. El Learning Analytics (LA) es la medición, recopilación, análisis y datos sobre los alumnos y sus contextos, con el propósito de comprender y optimizando el aprendizaje y los entornos en los que se produce para cubrir la mayoría de la investigación educativa, pero típicamente se combina con dos suposiciones: que el aprendizaje analítico hace uso de datos preexistentes, legibles por máquina, y que sus técnicas pueden ser usadas para manejar grandes datos, grandes conjuntos de datos que no serían factible tratar manualmente [7]. Learning Analytics es un espacio significativo de análisis del aprendizaje con tecnología que ha surgido durante los últimos 7 años [8]. El reporte horizon realizado por indica que el LA pretende utilizar el análisis de datos para generar información que permita tomar las mejores decisiones en ámbito educativo, para elaborar mejores pedagogías, entender a los estudiantes el porqué de su abandono de los estudios e incrementar la retención, está información ha sido eficaz y deben mantenerse; los resultados obtenidos son importantes para los directivos, los encargados de crear normativas y demás autoridades que son parte del sistema educativo. Para los docentes, el LA es crucial a la hora de buscar cómo interactúan los educandos con los textos y materiales disponibles por Internet. Los educandos también se benefician de los resultados de LA, mediante las diferentes aplicaciones desarrolladas para dispositivos móviles y plataformas por Internet que utilizan datos específicos de cada estudiante para crear sistemas de apoyo que se ajusten a las necesidades de aprendizaje. La selección de los métodos de análisis de datos similares entre el Análisis Estadístico Implicativo y el Learning Analytics se realizó utilizando los modelos y estándares (MSSS) [9] y la elaboración de un diseño cuasi-experimental en la ingeniería de software propuesto por Donald Campbell y Julian Stanley [10] , el objetivo principal es comparar el uso de memoria de las técnicas clustering LA (hclust.vector, dendro.variables y diana) y SIA (callHierarchyTree y callSimilarityTree) [11].

## II. Método

Por el paradigma de investigación es una investigación de tipo cuantitativo, por el tipo de diseño utilizado es quasi experimental, por el tiempo de estudio es transversal, el colectivo de estudio lo conforman las 100 000 bases de datos aleatorias formadas por 1000 observaciones y 100 variables, la población es la información sobre nombre del archivo, número de filas, número de columnas, total de datos, tiempo y memoria, por la amplitud de estudio es un muestreo de 383 bases de datos aleatorias binarias.

### *A. Materiales: el equipo informático y software*

Para el estudio se utilizaron tres computadores con el mismo microprocesador: Intel® Core ™ i7-CPU @ 2.2 Ghz y 8Gb de memoria RAM, se ha instalado los sistemas operativos Windows8-64 bits, Linux – Ubuntu 16.04-64 bits y MAC OS 10-64 bits. Todos los computadores y sistemas operativos trabajaron con el software estadístico libre R, versión 3.4.1; el entorno de desarrollo integrado libre RStudio, versión 1.0.143 y el paquete RCHIC, versión 0.24. Las bases de datos se generaron aleatoriamente utilizando la función runif() perteneciente al paquete estándar de R. Los datos utilizados fueron dicotómicos generados por la función runif() y round().

### *B. Metodología: diseño cuasi- experimental*

El tamaño de la población fue de 100000 bases de datos formadas hasta por un máximo de 1000 observaciones y 100 variables. Por su tamaño, se escogió una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media, se consideró la fórmula (1) para el cálculo de la muestra:

$$n = \frac{s^2}{\frac{E^2}{Z^2_{\alpha/2}} + \frac{s^2}{N}} \quad (1)$$

Para la aplicación de la fórmula se utilizaron los parámetros desviación estándar $s$ = 1; α = 5%; $Z$ = 1.96; $E$ = 10%; $N$ = 100000 y se generó un tamaño de la muestra de 382.675 que redondeado es 383. Las hipótesis estadísticas que se demostraron fueron normalidad según en test de Anderson-Darling, test de hipótesis de Kruskal-Wallis y su respectivo post test. Para demostrar las hipótesis se planteó un cuasiexperimento en la ingeniería de software de tipo $RGXO_1$. Donde RG representa el grupo aleatorio del grupo experimental (tanto-inter como intra-grupos), X representa el tratamiento que en este caso son los 3 técnicas cluster jerárquicos utilizadas en LA (hclust.vector, dendro.variables y diana) y 2 técnicas usadas en SIA ( callHierarchyTree y callSimilarityTree). Se trabajó con un nivel de significancia del 95%. La variable dependiente fue el espacio de memoria ocupado (en kilobytes) que es de tipo numérico.

## III. Análisis de Resultados

### *C. Análisis descriptivo*

Se procedió a realizar un gráfico de cajas y alambres comparativo para cada uno de los 5 métodos analizados, éste se muestra a continuación en la Fig. 1.

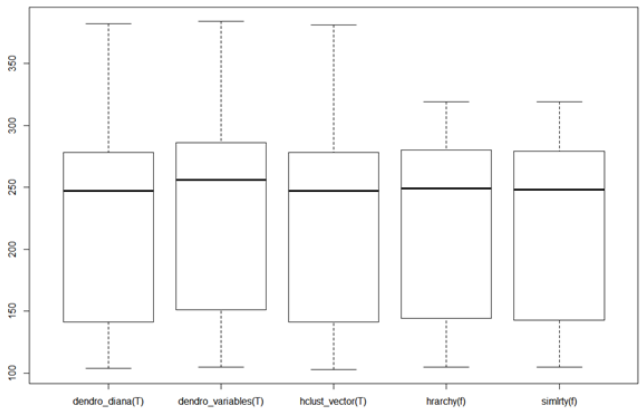


Fig. 1. Gráfico comparativo de cajas y alambres

La Tabla I, muestra un cuadro comparativo entre las medidas de centralización y el tamaño de muestra utilizado para cada uno de los métodos analizados.

TABLA I
CUADRO COMPARATIVO ENTRE MÉTODOS CLUSTER SU MEDIA Y EL TAMAÑO DE MUESTRA

| | dendro_diana(T) | dendro_variables(T) | hclust_vector(T) | hrarchy(f) | simlrty(f) |
|---|---|---|---|---|---|
| Centrali-zación | 222.9568 | 231.3444 | 223.0070 | 224.8100 | 223.8985 |
| Tamaño | 3447 | 3447 | 3447 | 3447 | 3447 |

### D. *Comprobación de supuestos*

Para determinar el test apropiado a utilizar se procedió a la comprobación de los supuestos. A continuación, se muestra la gráfica de cuartiles, que da una idea gráfica sobre la normalidad de los datos sobre la memoria ocupada por los diferentes métodos.

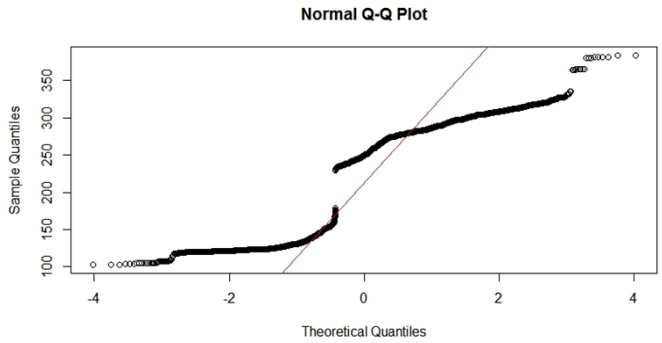


Fig. 2. Gráfico de cuartiles

$H_o$: No se observa diferencia entre los datos de ocupación de memoria y la distribución normal

$H_1$: Se observa diferencia entre los datos de ocupación de memoria y la distribución normal

Para su demostración se utilizó el test de Anderson-Darling, cuyos resultados se muestra en la Tabla II:

TABLA II
RESULTADOS DE LOS TEST DE NORMALIDAD

| TEST DE NORMALIDAD | ESTADÍSTICO | P-VALUE |
|---|---|---|
| Anderson-Darling | A = 1224 | < 2.2e-16 |

### E. *Prueba de hipótesis*

La hipótesis estadística por demostrar se muestra a continuación:

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4 = \tilde{\mu}_5$$
$$H_1 : \exists i, j \in \{1,2,3,4,5\} / \tilde{\mu}_i \neq \tilde{\mu}_j \quad (2)$$

Se utilizó el test de hipótesis no paramétrico suma de rangos para muestras independientes de Kruskal-Wallis, los resultados entregados por la función kruskal.test(x, y) del paquete estándar del software R fueron los siguientes :

```
                  dendro_diana(T) dendro_variables(T) hclust_vector(T)
dendro_variables(T) < 2e-16       -                   -
hclust_vector(T)    0.873         < 2e-16             -
hrarchy(f)          0.026         2.7e-12             0.036
simlrty(f)          0.491         < 2e-16             0.500
                  hrarchy(f)
dendro_variables(T) -
hclust_vector(T)    -
hrarchy(f)          -
simlrty(f)          0.500
```

con las salidas se elaboró la gráfica de la Fig. 3.

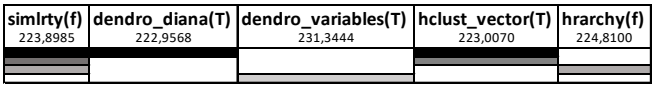


Fig. 3. Grupos de homogeneidad

## IV. DISCUSIÓN

El gráfico de cajas y alambres nos muestra una homogeneidad en la dispersión de los 3 métodos de Learning Analytics y también en los de las técnicas del Análisis estadístico Implicativo, pero curiosamente las técnicas SIA son más homogéneas. En cuanto a las medidas de centralización se puede observar que aparentemente la ocupación de memoria es similar entre los 5 métodos, con una aparente mayor ocupación del método dendro.variables. Antes de realizar la prueba de hipótesis, se procedió a comprobar sus supuestos. Se realizó la prueba de normalidad de Anderson-Darling que nos dio un p-valor de 2.2e-16 indicándonos que se debe rechazar la hipótesis nula y que por tanto los datos no han sido extraídos de una población normal, este resultado se corrobora con el gráfico de cuartiles que muestra un gran alejamiento de la distribución de datos a los cuartiles teóricos. El no cumplimiento de este supuesto es suficiente para optar por los test no paramétricos. El test no paramétrico seleccionado es el test de hipótesis no paramétrico de suma de rangos para muestras independientes de Kruskal-Wallis que nos entrega un valor de chi cuadrado y un p-valor de 2.2e-16, que nos indica con un alto valor de significancia que se rechaza la hipótesis nula y por lo tanto al menos un par de los 5 métodos clusters son diferentes. Con el objetivo de determinar la relación entre los pares se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes con el cual se ha construido el gráfico de homogeneidad de medidas de centralización que nos indica que hay cuatro grupos bien definidos, de los cuales nos llama la atención el método dendro.variables que es el que más ocupación de memoria tiene y además nos indica que la técnica callSimilarityTree del análisis estadístico Implicativo, dendro.variables y diana de Learning Analytics son los que menos ocupación de memoria tienen.

## V. CONCLUSIONES

El objetivo de esta investigación fue determinar si desde el punto de vista de la ocupación de memoria las técnicas utilizadas en el Análisis Estadístico Implicativo (SIA) son similares con los métodos cluster utilizados en Learning Analytics (LA). Con la técnica de modelos y estándares (MSSS) se determinó las técnicas de análisis similares entre SIA y LA. En SIA se consideraron la función que genera el árbol de similaridad (callSimilarityTree) y la función que genera el árbol de cohesión (callHierarchyTree), mientras que en LA se consideraron los siguientes tres métodos: hclust.vector, dendro.variables y diana. La hipótesis propuesta es que no existe diferencia significativa en la ocupación de memoria entre los 3 métodos de LA y las dos técnicas SIA. Se utilizó un test ANOVA no paramétrico y una posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para llegar a la conclusión de que desde el punto de vista de la ocupación de memoria: el método menos recomendable es dendro.variables dentro de LA por

ocupar más memoria, los métodos más recomendables son hclust_vector y dendro_diana de LA y simlrty dentro de SIA. Las relaciones completas en ocupación de memoria se muestran a continuación:

**simlrty** (en SIA) = **dendro_diana** (en LA)
= **hclust_vector** (en LA)
< **hrarchy** (en SIA)
< **dendro.variables** (en LA)

Es importante notar que el método callSimilarityTree dentro del Análisis Estadístico Implicativo ocupa el espacio mínimo de memoria al igual que los métodos cluster de LA: dendro.variables y diana, pero por investigaciones anteriores se demostró que las técnicas callSimilarityTree y callHierarchyTree son mucho más rápidas [11].

## REFERENCIAS


[1] R. Gras, "Panorama du développement de l'ASI à partir de situations fondatrices," *Actes des Troisièmes Rencontres Internationale ASI Analyse Statistique Implicative*, Volume Secondo supplemento al, pp. 9-33, 2005.

[2] Y. Le Bras13, P. Meyer13, P. Lenca13, and S. Lallich, "A robustness measure of association rules."

[3] R. Pazmiño, "Aproximación al Análisis Estadístico Implicativo desde sus Aplicaciones Educativas," 2014.

[4] R. A. Pazmiño-Maji, F. J. García-Peñalvo, and M. Á. Conde-González, "Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers," 2017.

[5] R. Couturier and R. Gras, "CHIC: traitement de données avec l'analyse implicative," in EGC, 2005, pp. 679-684.

[6] R. Pazmiño, F. J. García-Peñalvo, R. Coutrier, and M. Conde-González, "Statistical implicative analysis for educational data sets: 2 analysis with RCHIC," 2015.

[7] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal for the Scholarship of Teaching and Learning*, vol. 4, p. 17, 2010.

[8] R. Ferguson, "Learning analytics: drivers, developments and challenges," 2016.

[9] L. C. Briand, Y. Labiche, M. D. Penta, and H. Yan-Bondoc, "An experimental investigation of formality in UML-based development," *IEEE Transactions on Software Engineering*, vol. 31, pp. 833-849, 2005.

[10] D. T. Campbell and J. C. Stanley, Experimental and quasi-experimental designs for research, 1966.

[11] R. A. Pazmiño-Maji, F. J. García-Peñalvo, and M. A. Conde-González, "Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics," in *Proc. of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2017, p. 49.