Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

Análisis de clústeres para la clasificación de datos económicos

Rómulo Alejandro Barba López[1], Humberto Stanley Guerrero Herrera [2] Janeth Doris Salazar González[3]

1 Universidad de Guayaquil, romulo.barba@ug.edu.ec

2. Universidad de Guayaquil. humbertoguerreroh@hotmail.com

3. Universidad de Guayaquil. janethdsg@hotmail.com

RESUMEN

El artículo tuvo como objetivo aplicar la clasificación por clusters para agrupar los datos reportados en relación con los volúmenes de ventas y exportación reportados para las empresas en el Catálogo Central de Datos de Ecuador. Se decidió para ello el empleo del Rapidminer para poder recomendar la utilización de este paquete en posibles situaciones de clases. Se realizó el agrupamiento por clústeres empleando el algoritmo de k-medias, para 2536 empresas en relación con los datos reportados por estas en cuanto a; ventas totales, ventas nacionales, total de exportaciones netas, total de empleados, total de empleados hombres y total de empleados mujeres. Los operadores empleados y presentes en Rapidminer hicieron factible: la fácil lectura del fichero, el empleo del algoritmo de k medias y la obtención de tablas de correlaciones entre las variables señaladas. La posibilidad de representación gráfica facilitó un análisis desde multiples perspectivas.

En relación con el aprendizaje de Rapidminer debe señalarse que este paquete presenta todo un conjunto de operadores que no son de fácil utilización para personal no especializado en estadística y en minería de datos. No obstante si se enfoca el aprendizaje de este en relación con aspectos específicos, como se realizó para el agrupamiento en clústeres y el establecimiento de correlaciones, si pudiera analizarse su posible aplicación en la docencia de las ciencias administrativas. A le vez las posibilidades gráficas que ofrece esta herramienta son muy atractivas para la presentación de resultados.

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

Palabras claves: rapidminer, agrupamiento por clusters, k medias

Cluster analysis for the classification of economic data

ABSTRACT

The article aimed to apply the classification by clusters to a group data reported in relation to sales and export volumes in the Central Data Catalog of Ecuador. It was decided to do the use of RapidMiner to recommend this package in possible situations in class. Clustering was performed using k-means algorithm, for companies in relation to 2536 data reported by these in terms: total sales, national sales total exports, total employees, employees men and women total employees. Operators present in RapidMiner made possible: easy reading of the file, use of the k-means algorithm and obtaining correlation tables between the mentioned variables. The possibility of graphic representation provided an analysis from multiple perspectives.

Regarding learning RapidMiner should be noted that this package has a set of operators that are not easy to use for non-specialists in statistics and data mining. However if learning this in relation to specific aspects, as it was done for clustering and establishing correlations, if you could be possible its application in the teaching of administrative sciences. The graphic possibilities offered by this tool are very attractive to the presentation of results.

Keywords: Rapidminer, clustering, k-means

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

1. INTRODUCCIÓN

La clasificación de los datos económicos se inscribe dentro de la problemática general de las aportaciones de la Matemáticas a la metodología económica en que (Sánchez, 2000) señalaron el debate al respecto. La aplicación del análisis de clústeres para clasificar los datos de una matriz ha sido estudiada desde hace muchos años (Hartigan, 1972) y se han reportado aplicaciones en los estudios de mercado (Punj & Stewart, 1983) en la segmentación de estos (Sharma & Lambert, 2013) en el análisis del risgo financiero (Kou, Peng, &.Wang, 2014). En la práctica contable el empleo de diversas técnicas incluyendo la minería de datos ha ganado importancia creciente y Hernandez (2015).aplicó el análisis de clústeres para el agrupamiento de la data contable y el posible estudio de tendencias en esta.

En relación con las herramientas que se pueden utilizar para la clasificación en clústeres existen diferentes posibilidades bien empleando el software SPPS en sus diferentes versiones o como una de las posibilidades que ofrecen los paquetes de minería de datos como el Rapidminer (Jungermann, 2009). Este paquete ha sido empleado por Pazmiño Santacruz y González Alonso (2014) para la clasificación de las publicaciones relacionadas con la comunicación organizacional en Pymes-y estos autores concluyeron que el agrupamiento por clústeres utilizando Rapidminer (Rapidminer, 2014) permitió clasificar en cinco grupos los articulos analizados. González Alonso y González (2015) han utilizado ese mismo paquete para la clasificación por clústeres en este caso la agrupación por países de las revistas latinoamericanas y emplearon el algoritmo de k-medias con distancias euclidianas.

De acuerdo con estos resultados reportados en la literatura este artículo se trazó como objetivo aplicar la clasificación por clusters para agrupar los datos reportados en relación con los volúmenes de ventas y exportación reportados para las empresas en el Catálogo Central de Datos de Ecuador (INEC, 2015). Se decidió para ello el empleo del Rapidminer con el objetivo de poder recomendar la utilización de este paquete en posibles situaciones de clases.

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

2. METODOS

.La Data que se tomó fue la Base de Datos que se tomó fue la correspondiente a las Empresas reportadas en el Catalogo Central de Datos de Ecuador, para el año 2013 (INEC, 2015).

El Modelo que se empleó para realizar el agrupamiento de los datos se presenta en la Figura 1 y se estructuró en Rapidminer.

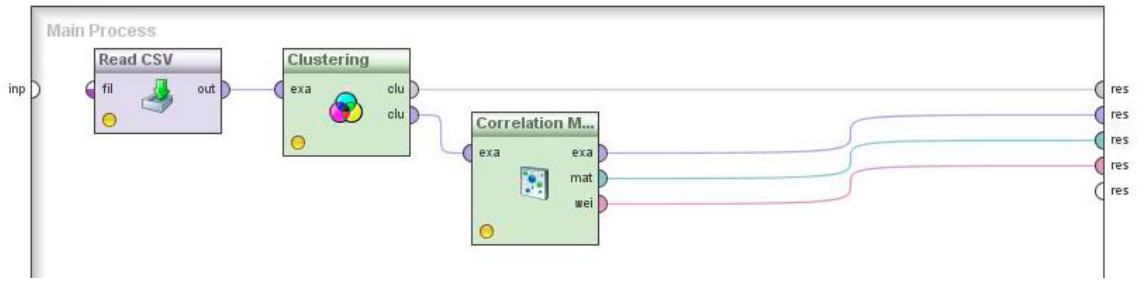


Fig. 1. Proceso en RapidMiner para el tratamiento de la data.

Como se observa el proceso consta de tres operadores esenciales:

La lectura del fichero en Excel (CSV). En este caso se el fichero estaba delimitado por comas.

Agrupamiento en Clústeres. En este caso se utilizó el algoritmo de k-medias que es uno de los más frecuentemente utilizados (Jain, 2010) . El Rapidminer ofrece distintos algoritmos para ello. En este caso mediante pruebas iniciales se ajustó el valor k= 5 y empleo de distancias ecuclideanas.

Establecimiento de Matriz de correlaciones. Uno de los aspectos más importantes del RapidMiner es que ofrece la posibilidad de otros análisis bien previos a la clasificación en clústeres, o después de esta. Como se observa en esta caso el establecimiento de matrices toma los datos del fichero inicial y la salida se obtiene tanto la clasificación por clústeres, como tres posibilidades de presentación de la matriz.

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

3. RESULTADOS

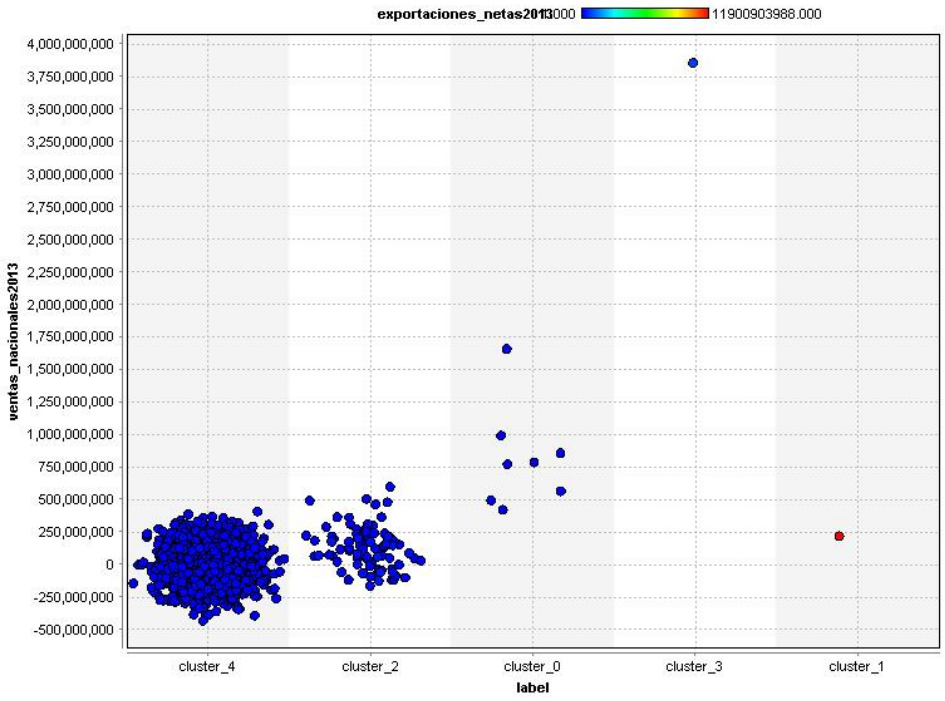
Una de las posibilidades del agrupamiento por clústeres es que permite clasificar con respecto a diferentes variables. En este caso se agrupo en relación con: exportaciones netas, ventas totales y ventas nacionales. Todas estas referidas al año 2103 que es el que ofrece la base de Datos analizada.

En la Tabla 1 se presentan los valores de los centroides calculados para ellas y que se obtienen directamente de la interfase de Rapidminer:

Tabla 1. Valores calculados para los centroides

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
exportaciones_netas2013	2.E+16	1.E+20	5.E+22	7.E+16	3.E+16
ventas2013	9.E+18	1.E+20	2.E+24	4.E+18	8.E+15
ventas_nacionales2013	9.E+17	6.E+14	1.E+24	4.E+18	6.E+14

En la Figura 2 se presenta el agrupamiento obtenido para las 2536 empresas que informaron estos datos



Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

Fig. 2. Agrupamiento en clústeres de 2536 empresas (exportaciones, ventas totales, ventas nacionales).

La posibilidad visual que ofrece RapidMiner es también interactiva por lo que permite que el investigador se concentre bien en un grupo o en una empresa especifica.

El agrupamiento por clústeres se realizó también incorporando más variables que fueron a más de las consideradas el total de empleados por empresas dividido además en hombres mujeres.

La Tabla 2 presenta las correlaciones obtenidas:

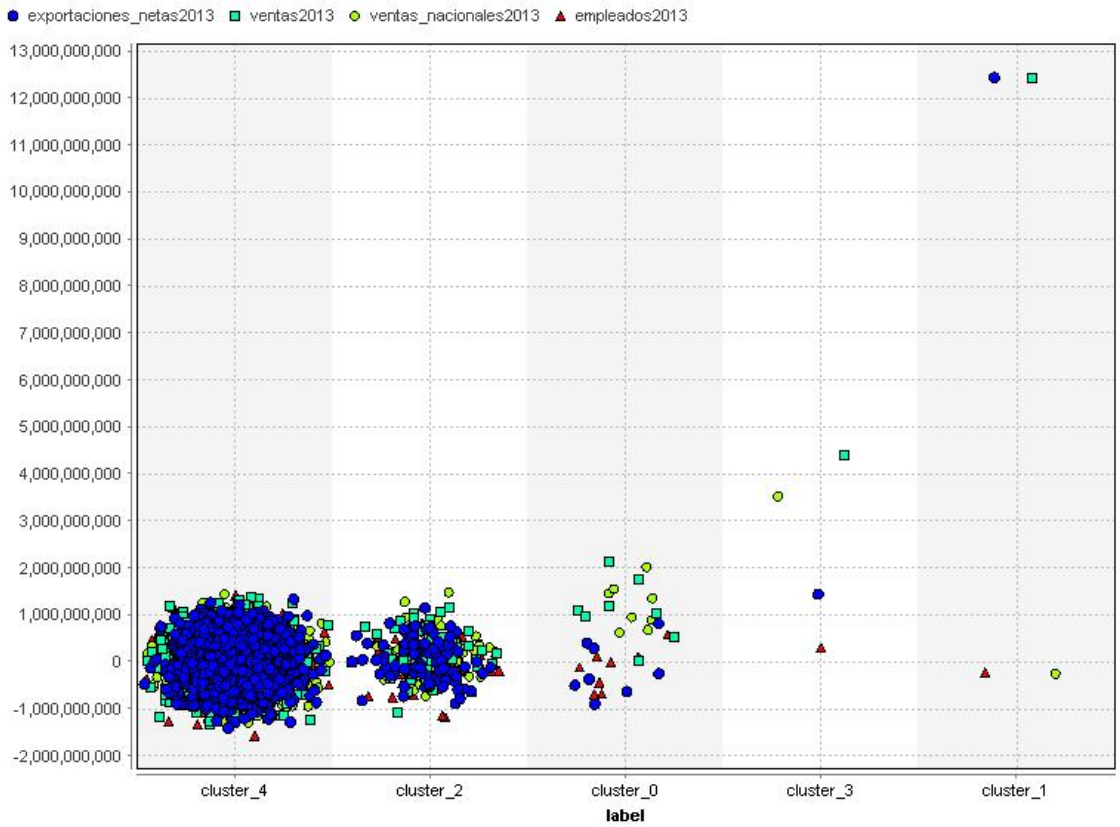
Tabla 2. Coeficiente de Correlación entre las variables consideradas para el año 2013

Primer Atributo	Segundo Atributo	Coef. Correlación
exportaciones_netas	Ventas	0.933
Ventas	Empleados	0.541
ventas_nacionales	emp_hombres	0.535
ventas_nacionales	emp_mujeres	0.433
exportaciones_netas	Empleados	0.270
exportaciones_netas	ventas_nacionales	0.062

En esta tabla se omitieron las correlaciones entre los totales de empleados y la distribución por sexos y las variables ya mencionadas (exportaciones netas, ventas nacionales y ventas totales), se observa que si existe una correlación alta (0.933) entre exportaciones y total de ventas, pero que estas aparecen débilmente correlacionadas con el total de empleados y mucho menos con las ventas nacionales

La Figura 3 a continuación ejemplifica la posibilidad de representación gráfica del agrupamiento por clusters, para las cuatro variables: exportaciones, ventas totales, ventas nacionales y total de empleados:

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304



El grafico representa claramente que estamos ante cinco grupos diferentes de empresas uno ,mayoritario (clúster 4) seguido del cluster 2 y tres grupos con otros resultados.

4. CONCLUSIONES

En relación con los objetivos trazados para la investigación se pudo concluir que se pudo realizar el agrupamiento por clústeres para un total de 2536 empresas en relación con los datos reportados por estas en cuanto a; ventas totales, ventas nacionales, total de exportaciones netas, total de empleados, total de empleados hombres y total de empleados mujeres. Los operadores empleados y presentes en Rapidminer hicieron factible: la fácil lectura del fichero, el empleo del algoritmo de k medias y la obtención de tablas de correlaciones entre la variables señaladas. La posibilidad de representación gráfica de acuerdo con el agrupamiento obtenido por clústeres en rlación con un grupo de variables facilita un análisis desde multiples perspectivas.

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

En relación con el aprendizaje de Rapidminer debe señalarse que este paquete presenta todo un conjunto de operadores que hacen que no resulta simple su utilización para personal no especializado tanto en los aspectos estadísticos, como en minería de datos. No obstante si se enfoca el aprendizaje de este en relación con aspectos específicos, como se realizó para el agrupamiento en clústeres y el establecimiento de correlaciones, si pudiera analizarse su posible aplicación en la docencia de las ciencias administrativas. A le vez las posibilidades gráficas que ofrece esta herramienta son muy atractivas para la presentación de resultados.

5. REFERENCIAS BIBLIOGRÁFICAS

González Alonso, J. A., et al.González, Y. P. (2015). Análisis de las revistas latinoamericanas de acceso abierto. El caso ecuador. *Revista Publicando, 2*(2), 12-23.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association, 67*(337), 123-129.

Hernandez, A. B. (2015). La detección del fraude contable utilizando técnicas de minería de datos. *Revista Publicando, 2*(5), 103-113.

INEC. (2015). Estadisticas. *Estadisticas Agropecuarias.* Retrieved Febrero, 2016, de http://www.ecuadorencifras.gob.ec/estadisticas-agropecuarias-2/

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters, 31*(8), 651-666.

Jungermann, F. (2009). *Information extraction with rapidminer*. Paper presented at the Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities.

Kou, G., Peng, Y., et al.Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using mcdm methods. *Information Sciences, 275*, 1-12.

Pazmiño Santacruz, M. R., et al.González Alonso, J. A. (2014). Análisis exploratorio sobre las publicaciones relacionadas con la comunicación organizacional en pymes. *Revista Publicando, 1*(1).

Revista Publicando, 3(7). 2016, 267-275. ISSN 1390-9304

Punj, G., et al.Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 134-148.

Sánchez, Á. C. (2000). Aportaciones de la matemática a la metodología económica. *Psicothema*, *12*(2), 103-107.

Sharma, A., et al.Lambert, D. M. (2013). Segmentation of markets based on customer service. *International Journal of Physical Distribution & Logistics Management*.