Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Una Introducción al Análisis Cluster.

Lucero Martínez Bonilla[a], Fernando Velasco Luna[b], Francisco S. Tajonar Sanabria[c].
[a,b,c] *Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Puebla, Puebla, México.*
[a] fvelasco@fcfm.buap.mx, [b] fvelasco@fcfm.buap.mx, [c] ftajonar@fcfm.buap.mx.

Resumen

La técnica de análisis cluster consiste en clasificar elementos formando grupos conglomerados o cluster de a los individuos en estudio, de tal forma que los individuos dentro de cada cluster presenten cierto grado de homogeneidad en base a los valores adoptados sobre un conjunto de variables. En esta técnica los conglomerados son desconocidos y el proceso consiste en su formación de tal manera que las unidades en cada cluster sean homogéneas. En este trabajo se presentan algunos trabajos realizados por autores en el cual se usa en el análisis cluster. Se menciona el resumen y algunos párrafos de los artículos hechos por los autores. En trabajos futuros se pretende abordar la parte teórica de la técnica de análisis cluster.

Palabra claves: Análisis Mutivariado, Conglomerados, Variables.

Introducción

La técnica de análisis cluster consiste en clasificar elementos formando grupos conglomerados o cluster de a los individuos en estudio, de tal forma que los individuos dentro de cada cluster presenten cierto grado de homogeneidad en base a los valores adoptados sobre un conjunto de variables. En esta técnica los conglomerados son desconocidos y el proceso consiste en su formación de tal manera que las unidades en cada cluster sean homogéneas. Es decir, el Análisis Cluster, conocido también como Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

Recordar que en el analisis multivariado cada individuo depende de un gran número de variables, así los clusters formados vendrán determinados por las múltiples variables usadas en el estudio. En el análisis cluster se definen grupos tan distintos como sea posible en función de los datos.

Por ejemplo, puede clasificarse un grupo de trabajadores de un municipio según ciertas características personales y socioeconómicas (salario, sexo, nivel cultural, etc) lo que proporciona una segmentación del municipio.

Algunos trabajos

Grisales Romero (2006), menciona que "Los métodos estadísticos multivariados pueden agruparse en dos conjuntos: los que permiten extraer la información acerca de la interdependencia entre las variables que caracterizan cada uno de los individuos, y los que permiten extraer información acerca de la dependencia entre una (o varias) variable(s) con

otra (u otras). Del primer grupo referido hacen parte el análisis factorial, de clusters, de correlación canónica, de ordenamiento multidimensional y de correspondencias; en el segundo grupo se encuentran las técnicas de regresión, los análisis de contingencia múltiple, el análisis de la covarianza y el análisis discriminante."

El analisis cluster es un método estadístico multivariante de clasificación de datos. A partir de una tabla de casos-variables, trata de situar los individuos en grupos homogéneos, conglomerados o clusters, de manera que individuos que puedan ser considerados similares sean asignados a un mismo conglomerado, mientras que individuos diferentes se localicen en clusters distintos.

La creación de grupos basados en *similaridad* de casos exige una definición de este concepto, o de su complementario *distancia* entre individuos.

La variedad de formas de medir diferencias multivariables o *distancias* entre casos proporciona diversas posibilidades de análisis. El empleo de ellas, y el de las que continuamente siguen apareciendo, así como de los algoritmos de clasificación, o diferentes reglas matemáticas para asignar los individuos a distintos grupos, depende del fenómeno estudiado y del conocimiento previo de posible agrupamiento que de él se tenga.

Puesto que la utilización del análisis cluster ya implica un desconocimiento o conocimiento incompleto de la clasificación de los datos, el investigador ha de ser consciente de la necesidad de emplear varios métodos, ninguno de ellos incuestionable, con el fin de contrastar los resultados.

Existen dos grandes tipos de análisis de clusters: ***no jerárquicos*** y ***jerárquicos***.

Se conocen como ***no jerárquicos*** a aquellos que asignan los casos o grupos diferenciados que el propio análisis configura, sin que unos dependan de otros. Se denominan ***jerárquicos*** a los que configuran grupos con estructura arborescente, de forma que clusters de niveles más bajos van siendo englobados en otros clusters de niveles superiores.

Una vez finalizado un análisis de clusters, el investigador dispondrá de una colección de casos agrupada en subconjuntos jerárquicos o no jerárquicos. Podrá aplicar técnicas estadísticas comparativas convencionales siempre que lo permita la relevancia práctica de los grupos creados; así como otras pruebas multivariantes, para las que ya contará con una variable dependiente *grupo*, aunque haya sido creada artificialmente.

El análisis cluster se puede utilizar para agrupar individuos (casos) y también para agrupar variables. En adelante, cuando se hace una referencia a grupos de individuos (o casos) debe sobreentenderse que también se hace a un conjunto de variables. El proceso es idéntico tanto si se agrupan individuos como variables.

Los llamados métodos jerárquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Nuñez y Escobedo (2011) mencionan "Como primer punto para las futuras explicaciones en este ensayo sobre el análisis clúster, se definirá el concepto de Unidad Básica de Caracterización (UBC) que es definida como la unidad básica del objeto (planta, animal o entidad) que se va ha describir y que a su vez será empleada para lograr el objetivo de la caracterización; dependiendo del propósito del estudio de caracterización, la UBC

puede ser un individuo, una población silvestre, una línea o un híbrido. (González-Andrés 2001).” y “El análisis clúster es un método matemático que está incluido en lo que hoy se llama estadística multivariada o estadística multivariante; este método es principalmente utilizado para la formación de grupos de UBC’s con características similares a partir de las similitudes o disimilitudes que se presentan entre pares de estas UBC’s en *n* características evaluadas (Johnson 1998). Este tipo de análisis está compuesto por dos métodos interrelacionados e igualmente importantes. El primero es el cálculo de los índices de similitud o de disimilitud entre pares de UBC’s, del cual existe un sin número de referencias en la literatura; no obstante, estos índices deben ser aplicados de acuerdo a la naturaleza de los datos y al objetivo de la caracterización. Y el segundo es la aplicación del método de aglomeración adecuado, donde también existe mucha literatura al respecto, que permite a partir de los índices de similitud o disimilitud generar las gráficas de árbol o dendrogramas que son representaciones gráficas donde el investigador puede tener de una manera resumida el parecido que presentan los grupos de UBC’s. El método de aglomeración a utilizar principalmente está definido por el objetivo de la caracterización. Los métodos, cálculo de los índices de similitud y la aplicación del método de aglomeración, englobados en el análisis clúster, a pesar de que presentan unas sólidas bases matemáticas no son tan restrictivos respecto a sus bases estadísticas (Johnson 1998). Por lo que el principal problema que presenta este análisis es la elección del índice de similitud o disimilitud y del método de aglomeración más apropiados; la elección de la mejor combinación de los métodos del análisis clúster depende principalmente de la naturaleza de los datos (si son datos doble estado, multi-estado con o sin secuencia lógica, cuantitativos, genéticos o secuencias de ADN/proteínas) y el objetivo de la caracterización. Caracterizar es un método no es un objetivo.”

García, Blázquez y López (2012), realizan un estudio empírico utilizando la técnica de análisis cluster sobre la capacidad de innovación tecnológica en Latinoamérica y España. Con apoyo en la literatura sobre capacidad nacional de innovación y economía del cambio tecnológico, se realizó un estudio empírico utilizando la técnica de análisis estadístico multivariante de *cluster* y los indicadores de innovación tecnológica publicados en el *Global Competitiveness Report 2002-2003* y *2009-2010* (WEF, 2002; WEF, 2009), para explorar la existencia de grupos de países caracterizados por diferentes niveles de innovación tecnológica, profundizando en sus características y en la distancia que les separa, así como en su evolución a lo largo del período 2002-2009. Los resultados muestran la existencia de cuatro grupos de países que se definen por una distinta capacidad de innovación tecnológica, tanto en lo referente a política tecnológica gubernamental y empresarial como en lo relativo a infraestructuras tecnológicas y capital humano. Los grupos también difieren en su evolución en relación con estos factores en el período considerado.

Castro y Cols. (2012) presentan con este estudio la aplicación del Análisis Clúster (AC) como método exploratorio de datos en registros múltiples de información pluviométrica. Se empleó el análisis multivariado en 150 estaciones de medición de precipitación mensual localizadas en el departamento del Valle del Cauca, Colombia. Se utilizaron las técnicas de Encadenamiento Simple, Ward y Centroide como métodos jerárquicos de aglomeración y la Distancia Euclídea al Cuadrado (DEC) como medida de similitud. El objetivo principal del estudio consistió en comprobar la hipótesis que las estaciones atípicas, es decir, aquellas que el AC agrupa individualmente (cambio en la varianza y la media), son de tipo no homogéneo. Se utilizó un análisis exploratorio gráfico y cuantitativo con series univariadas para comprobar dicha

hipótesis. Los resultados mostraron que mediante el AC se pueden obtener las estaciones no homogéneas, como también las estaciones cuyo comportamiento no es representativo de la muestra, dado que los grupos formados por esta técnica tienden a contener elementos muy parecidos entre sí, como los de máxima homogeneidad, excluyendo los que no pertenecen a esta clasificación.

El trabajo expone algunas aplicaciones del método cluster a la solución de problemas que se presentan en los talleres y secciones de tratamiento térmico, enfatizándose en las cuestiones relaciona das con el agrupamiento de piezas y la formación de células flexibles de fabricación, lo cual constituye la base para la introducción de todo un conjunto de técnicas y criterios de avanzada en la actividad pro ductiva de estas instalaciones.

El análisis de cluster es una técnica estadística que trata de identificar grupos de objetos o casos similares, basados en las propiedades de sus atributos. , en los últimos años se han reportado múltiples aplicaciones de este método en la industria y los servicios en general.

Ferreira., Rial y Varela (2010) ilustran las ventajas asociadas a la aplicación de técnicas multivariadas, el Análisis Cluster en dos fases, realizando una segmentación a posteriori (post hoc) basada en las preferencias de los turistas españoles. Ellos mencionan "Es innegable la importancia del turismo en la economía de países como Portugal, España y Francia, concretamente, en el Producto Interno Bruto y en la creación de empleo. Asimismo, el turismo en estos países potencia un crecimiento sustentable equilibrado. Así, la perspectiva del Marketing Turístico surge como un paradigma compartido por las potencias turísticas internacionales. Establecer una política de Marketing Turístico presupone conocer las necesidades y preferencias de los turistas, de modo de orientar de forma óptima las estrategias llevadas a cabo por los gobiernos y organizaciones responsables de la gestión de recursos turísticos, sean ellos naturales, humanos o infraestructurales. El principal objetivo del presente estudio es ilustrar las ventajas asociadas a la aplicación de técnicas multivariadas, el Análisis Cluster en dos fases, realizando una segmentación a posteriori (post hoc) basada en las preferencias de los turistas españoles. Los resultados obtenidos permitieron identificar 5 clusters: los Culture Seekers, que se caracterizan por preferir destinos con una alta oferta cultural; los Culture Seekers oriented by Low Prices, turistas que prefieren destinos con elevada oferta cultural y precios accesibles; los Sun and Tranquility Seekers, que son turistas que buscan destinos de sol, con ambientes tranquilos; los Sun and Night Lovers, referente a los turistas que prefieren destinos de sol, con elevada oferta de diversión nocturna; y los Night Lovers oriented by Low Prices, que son turistas que buscan destinos con elevada oferta de diversión nocturna y precios accesibles."

Borracci y Arribalzaga (2005), en su trabajo mencionan "Los programas educacionales han **r**ecurrido a distintos modelos de regresión lineal múltiple, de selección asistida por computadora y más recientemente, de redes neuronales artificiales para la confección de listados preliminares de mérito entre los postulantes a la residencia. El objetivo del trabajo fue evaluar y rediseñar un sistema para la selección y clasificación de aspirantes a un programa de residencias universitarias por medio de la aplicación de modelos de análisis multivariante y de redes neuronales artificiales. Material y Método: El diseño consistió en un estudio retrospectivo-transversal, realizado en un hospital universitario. Se evaluó una muestra al azar de 213 aspirantes a un programa de residencias médicas universitarias teniendo en cuenta el promedio de la carrera de grado, el resultado del examen de ingreso a la residencia, los antecedentes curriculares y biográficos, el internado y el puntaje de las entrevistas. Se

aplicó un análisis de conglomerados jerárquico (clúster análisis) para la clasificación y selección de los candidatos en un orden de mérito en base a los puntajes estandarizados de las 5 variables. Resultados: El análisis de conglomerados jerárquico clasificó 209 aspirantes en 12 conglomerados en base al promedio estandarizado de los valores obtenidos de las 5 variables. Este análisis se usó para construir una clasificación descriptiva de los grupos y una lista final por mérito de acuerdo a la posición relativa de cada candidato por encima o debajo de los puntajes promedios. Se imitó la solución de conglomerados por medio de una red perceptrón multicapa con una sensibilidad y especificidad de 94.1 y 99.1% respectivamente. Conclusiones: El análisis de conglomerados jerárquico fue un método útil y novedoso para clasificar una muestra de aspirantes a la residencia en conglomerados de acuerdo a la posición relativa de sus puntajes estandarizados por encima o por debajo de la media de todo el conjunto. Además, se entrenó una red PMC que permitió imitar los resultados del análisis de conglomerados con la suficiente precisión como para ser considerado un método alternativo de selección asistida por computadora cuando se trabaja con datos masivos. La solución de conglomerados constituye una aproximación alternativa para la selección de candidatos a la residencia.

Referencias

Borracci R.A. y Arribalzaga E.B. (2005), Aplicación de análisis de conglomerados y redes neuronales artificiales para la clasificación y selección de candidatos a residencias médicas. **Educación Médica,** 8(1): 22-30.

Castro Heredia, Lina M.; Carvajal Escobar, Yesid; Ávila Díaz, Álvaro Javier (2012) Analisis Cluster como técnica de Analisis Exploratorio de Registros Múltiples en Datos Meteorológicos. **Ingeniería de Recursos Naturales y del Ambiente**, núm. 11, pp. 11-20

Ferreira L.S.D., Rial B.A. y Varela M.J. (2010) Segmentación Post Hoc del Mercado Turístico Español. Aplicación del análisis cluster en dos etapas. **Estudios y Perspectivas en Turismo** Volumen 19, pp 592 – 606

García O.M.M., Blázquez M. L; y López S: J.I. (2012) Uso y aplicación de la técnica de análisis estadístico multivariante de cluster sobre la capacidad de innovación tecnológica en Latinoamérica y España **INNOVAR. Revista de Ciencias Administrativas y Sociales**, vol. 22, núm. 44, pp. 21- 39

Grisales, R.H. (2006) Usos y Limitaciones de los Métodos de Análisis Multivariados en a Investigación Epidemiológica. **Investigaciones Andina**, vol. 8, núm. 13, pp. 81-84.

Kohei Adachi (2018), **Matrix-Based Introduction to Multivariate Data Analysis,** Springer.

Núñez C.C y Escobedo, L, D. (2011), Uso correcto del análisis clúster en la caracterización de germoplasma vegetal. **Agronomía Mesoamericana**, vol. 22, núm. 2, pp. 415-427.