Incorporar actividades virtuales en educación superior: Algoritmo de Segmentación de docentes según sus competencias

Lucia Rosario Malbernat
Departamento de Sistemas, Universidad CAECE, Subsede Mar del Plata
Gascón 2464, Mar del Plata, Buenos Aires, República Argentina
+54 233 499-3400
lmalbernat@ucaecemdp.edu.ar; lmalbernat@gmail.com

Resumen

Para incorporar actividades virtuales en las carreras de grado, los docentes de las universidades deben innovar en sus prácticas docentes y para ello deben desarrollar competencias vinculadas con su preparación y actitud para la virtualidad.

En este trabajo se propone un algoritmo de segmentación, basado en el método del centroide o *k-means*[1], que agrupa a los docentes según su actitud innovadora tomando en consideración sus respectivas preparaciones y actitud para la virtualidad.

Se toman como variables de entrada la Preparación (índice P) y la Actitud (índice Q), -valores a los que se arriba mediante el cálculo de indicadores diseñados ad hoc- y se segmenta a los docentes identificando grupos o clústeres homogéneos con respecto a su vocación innovadora, clasificándolos en Innovadores, Indiferentes y Refractarios.

La información a la que se arribe con el análisis de los datos que surgen de la segmentación propuesta puede reducir la incertidumbre, por ejemplo, en relación a la toma de decisiones vinculadas con la selección de docentes, la incorporación de actividades online en las materias y la capacitación docente.

[1] Se ha tomado la adaptación de Hartigan y Worng (1979, pp. 100-108) del *k-means Clustering Algorithm*, publicado inicialmente por J.B. MacQueen en 1967.

Palabras clave: data mining; segmentación; innovación universitaria; TIC, educación virtual

Contexto

Se toma como caso de estudio a la Universidad CAECE Mar del Plata, República Argentina, en una investigación llevada a cabo sobre carreras de grado en el marco del cursado de la Maestría en Gestión Universitaria en la Universidad Nacional de Mar del Plata, la que diera lugar a la presentación del Informe de Tesis aprobado en noviembre de 2012.

Introducción

Se han tomado como variables de segmentación a los índices P y Q, cuantificados para cada docente en el marco de la investigación y se definieron las 3 categorías o clústeres (Innovadores, Indiferentes y Refractarios), identificadas a los fines del agrupamiento como A, B y C respectivamente, pues los profesores pueden clasificarse en, al menos, tres categorías [8], quienes generalmente tienen una actitud positiva hacia el uso de las TIC, alientan a sus estudiantes a adquirir conocimientos computacionales y por lo tanto aumentan los estándares de la enseñanza y el aprendizaje en todo el sistema, quienes asumen una posición neutral con relación al uso de las TIC en la educación y quienes tienen actitudes negativas explícitas hacia todas las nuevas tecnologías.

El modelo matemático diseñado ad hoc para cuantificar la preparación utiliza los indicadores nivel de uso de TIC, formación y experiencia en educación virtual y dominio de herramientas informáticas mientras que, para calcular la actitud para la virtualización, se entendieron necesarios los indicadores nivel de interés en el uso de TIC, interés en formación virtual, valoración del vínculo con las TIC y valoración a la educación virtual[2].

La metodología de segmentación descripta en este trabajo está basada en el uso de heurísticas que proporcionan una solución aproximada que se pretende buena para esta situación, que puede encontrarse en tiempo y a costo razonables, que mejorará el proceso de toma de decisiones reduciendo el nivel de incertidumbre. Es una técnica estadística bivariada, propia del Data Mining, cuya finalidad es segmentar, dividir un conjunto de elementos en grupos de modo que las características de sus elementos sean muy similares entre sí, con fuerte cohesión interna y sean disímiles intragrupos.

Dado que cada segmento debía agrupar docentes con características similares fue necesario elegir una medida para evaluar diferencias y similitudes. Una forma de medir la similitud es calcular la distancia entre pares de docentes. Por eso se tomaron los indicadores P (Preparación) y Q (Actitud) calculados a partir las elecciones hechas por los docentes al responder el cuestionario diseñado ad hoc. Una distancia reducida implicará mayor similitud que una distancia más amplia.

A partir del análisis del contexto de segmentación y de las características del caso, se prefirió utilizar un método no jerárquico, cuyo algoritmo particiona a partir de un elemento central de cada clúster o segmento, capaz de conglomerar a los restantes elementos del grupo a partir de mínimas distancias, denominado método de centroide o *k-means*, dónde k es un parámetro que define el número de elementos centrales o centroides (medias representativas de cada segmento) determinado por la cantidad de grupos o clústeres en que se desea segmentar (k coincide con el número de segmentos).

El objetivo de este método no es encontrar un grupo único y definitivo, sino ayudar a que el investigador obtenga una comprensión cualitativa y cuantitativa de los datos de modo de poder obtener grupos razonablemente similares [6].

En algunos contextos de segmentación se cuenta con datos de entrenamiento para diseñar el modelo, los cuales presentan un valor para la variable objetivo, es decir, los elementos a clasificar, ya están clasificados [9]. Para estos contextos, son apropiados los sistemas de clasificación supervisados que proponen el diseño de modelos a partir de los datos de entrenamiento. Para el caso de estudio, por el contrario, es apropiada una clasificación no supervisada.

Por otra parte, el número de clústeres incluidos en la segmentación puede ser o bien desconocido, o bien, conocido o dado por parámetro. Los métodos propuestos por muchos investigadores asumen esta última situación contextual [6], [2], [1], [3], [5], coincidente con el caso de estudio, en el que se han seleccionado 3 grupos.

Algunos autores [6], [1], [3], [10] proponen la elección al azar de los centros iniciales y otros proponen puntos iniciales depurados [2], [4], tal como ocurre en la presente propuesta en la que se conoce qué características se consideran buenas para los resultados pues se está en presencia de un agrupamiento con información externa [9].

Así, tratándose de un agrupamiento para el que se conocen de antemano las características de cada clúster y el rango de valores que pueden tomar las variables -datos calculados a partir del modelo matemático diseñado *ad hoc* que no presentarán valores extremos (*outliers*) que

[2] Ver Informe de Tesis "Innovación en educación universitaria: Factibilidad de incorporar actividades virtuales según las competencias docentes", 2012.

podrían dispersar los objetos del clúster-, se seleccionaron de manera sistemática los centroides iniciales, tomando 3 puntos equidistantes entre sí y de los límite superior e inferior de valores válidos.

Desarrollo

El algoritmo "*k-means*" encuentra una categorización que representa un valor óptimo según el criterio elegido [2], asignando a cada elemento el clúster del centroide más próximo siguiendo el procedimiento que se describe a continuación:

- Seleccionar k clúster iniciales ∈ δ, conjunto de clústeres. En el caso de estudio, k = 3 y δ = {A, B, C}
- Identificar casos (elementos) con valores centrales para definirlos como centroides iniciales de cada segmento. Los centroides iniciales, en el caso de estudio han sido definidos como $A_{(0)}$, $B_{(0)}$ y $C_{(0)}$.
- Repetir los siguientes pasos hasta que no se produzcan cambios significativos y no existan elementos equidistantes a 2 o más centroides.
  - Calcular las distancias Z de cada elemento a los 3 centroides iniciales.
  - Clasificar a cada elemento en el grupo del centroide más cercano (con menor valor de distancia).
  - Re-calcular los clúster iniciales promediando las variables de segmentación de cada clúster, es decir, obteniendo las medias de cada agrupación.
  - Volver a clasificar los elementos asignándolo al clúster del centroide más cercano.

Formalmente, se puede definir el modelo de segmentación que se propone de la siguiente manera para ω, conjunto de docentes: Sea k, cantidad de segmentos en los que se ha decido clasificar a los elementos w ∈ ω y δ, conjunto de clústeres, $S_i \subseteq \omega$, de la forma δ = $\{S_1, \ldots, S_k\}$, se cumplen las siguientes condiciones:

a. $S_i \cap S_i' = \varnothing$ con i ≠ i'.

b. $\bigcup_{i=1}^{k} S_i = \omega$

La primera condición establece que, dados dos segmentos S_i, S_i', no pueden tener elementos comunes (un docente no puede estar asignado a más de un grupo) y la segunda, que la unión de todos los segmentos S_i permite obtener al conjunto ω de docentes y que, por lo tanto, todo elemento w ∈ ω debe ser asignado a un grupo (todo docente debe ser situado en un segmento).

En consecuencia, se verifica que J, cantidad de elementos de ω coincide con la sumatoria de los j_i, cantidades de elementos de los segmentos S_i.

Ecuación 1 – Cálculo de J, cantidad de elementos de ω

$$\sum_{i=1}^{k} j_i = J ;$$

con J, cantidad de elementos de ω, k cantidad de segmentos y j_i cantidad de elementos del segmento S_i

El método debe encontrar una *k*-partición δ = $\{S_1, S_2, S_3\}$, dónde se maximice las similitud de los j_i elementos de una partición con respecto a los índices P y Q calculados para cada docente.

La maximización de la similitud de los elementos se ha logrado obteniendo las mínimas distancias Euclídeas[3] al cuadrado, es decir, mediante la suma de los cuadrados de las diferencias de los índices de cada elemento a clasificar y de los centroides definidos. Dicha distancia, expresada como $Z(x_w, y_l)$, con x_w un par ordenado (P_x, Q_x) que representa al elemento w a clasificar, el cual describe la preparación y actitud del docente e y_l el par ordenado (P_y, Q_y) que representa al Centroide l de un

[3] La distancia Euclídea tradicional calcula la longitud de la recta que une puntos en el espacio euclídeo: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

segmento S_i, es calculada con la siguiente función de distancia de x_w a y_l:

Ecuación 2 – Función de distancia $Z(x_w, y_l)$

$$Z(x_w, y_l) = |P_x - P_y|^2 + |Qx - Qy|^2 .$$

con P_x valor asignado al índice Preparación del elemento (docente) a clasificar, P_y el valor asignado al índice Preparación del centroide respecto del cual se va a calcular la distancia, y Q_x, Q_y, valores equivalentes correspondientes al índice Actitud.

Se cumplen para la función $Z(x_w, y_l)$ las siguientes propiedades que generalizan en geometría la noción de distancia entre 2 puntos [1]:

a) $Z(x_w, y_l) \geq 0$
b) ∀ w, $Z(x_w, x_w) = 0$, la distancia entre un elemento y sí mismo es cero;
c) $Z(x_w, y_l) = Z(y_l, x_w)$, la distancia es simétrica;
d) $Z(x_w, y_l) \leq Z(x_w, x_n) + Z(x_n, y_l)$, la distancia verifica la propiedad triangular.

La complejidad computacional del algoritmo *K- means* propuesto es lineal y, por lo tanto, eficiente. Se puede definir como O(2Jki) con J cantidad de docentes, k cantidad de segmentos e i, número de iteraciones; el 2 representa la cantidad de variables sobre las que se calcula la distancia $Z(x_w, y_l)$.

Sea el centroide de un clúster un elemento de la forma $Y_l = (P_y; Q_y)$, en el caso de estudio se han tomado para los grupos A, B, y C respectivamente, los siguientes centroides iniciales:

$A_{(0)} = (7{,}5; 7{,}5)$

$B_{(0)} = (5, 5)$

$C_{(0)} = (2{,}5; 2{,}5)$

Aplicando la función Z, la distancia de un elemento a cada centroide se calculó de la siguiente manera, dónde Q_x representa el valor Q (Actitud) del docente x y P_i, a su valor P (Preparación):

$Z(x_w; A_{(0)}) = |P_x - 7{,}5|^2 + |(Q_x - 7{,}5)^2$

$Z(x_w; B_{(0)}) = |P_x - 5|^2 + |(Q_x - 5)^2$

$Z(x_w; C_{(0)}) = |P_x - 2{,}5|^2 + |(Q_x - 2{,}5)^2$

Obtenidas las distancias de cada docente a cada centroide, -representado por su par ordenado (P_x; Q_x)-, se clasificó al docente asignándole la categoría más cercana (con menor valor de distancia).

El algoritmo básico *K-Means* propone [5], [4] calcular las medias de las distancias de los elementos del clúster y obtener así nuevos puntos centrales refinados.

Con los nuevos pares ordenados $A_{(1)}$, $B_{(1)}$ y $C_{(1)}$ se debe calcular nuevamente la asignación de categoría de cada caso provisionalmente clasificado. En consecuencia, se redefine al centroide l del segmento S_i que contiene j_i elementos como el promedio de las distancias de cada elemento del segmento al centroide l:

Ecuación 3 – Re-cálculo del Centroide l de Segmento S_i

$$l = \frac{\sum Z(x_w, y_l)}{j_i}.$$

Con x_w cada uno de los pares ordenados (P_x, Q_x) que representan elementos $w \in S_i$, y_l par ordenado que representa al centroide l que se re-calcula.

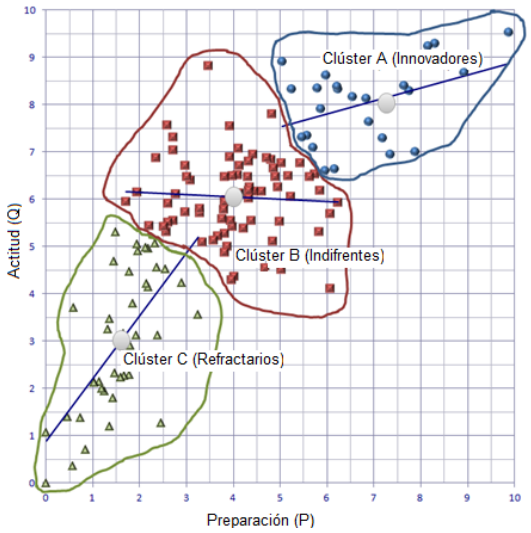
Este proceso de re-calcular los centroides tomando el promedio de las distancias de los puntos del segmento, re-calcular las distancias de los elementos y reasignar los elementos a un grupo según la distancia del elemento al centroide, se debe repetir hasta que no se produzcan clasificaciones dudosas y se puedan dar por clasificados a todos los docentes.

Resultados y objetivos

Surge de la aplicación del algoritmo que el 17,39% del total de la muestra, -24 docentes-, fue incluido en el clúster de los innovadores, la amplia mayoría del 53,62 % cayó en el segmento de Indiferentes y el 28,99% en el de Refractarios.

El Gráfico 1 muestra la clasificación final de cada sujeto de la muestra según los valores del par ordenado (P; Q) que lo que califican según lo indicado precedentemente y lo ubican en uno de los 3 segmentos definidos.

Gráfico 1- Segmentación docente



Con la información generada a partir de la aplicación del algoritmo no sólo se podrá reducir la incertidumbre al momento de diseñar un plan de capacitación docente. También se podrá observar la situación de cada carrera en relación a la factibilidad de incorporar actividades virtuales por contar ya con docentes preparados y con actitud positiva para hacerlo, pudiéndose en consecuencia, mejorar el proceso de toma de decisiones.

Formación de Recursos humanos

En noviembre de 2012, Lucía Rosario Malbernat obtuvo el título de Magister en Gestión Universitaria que expide la Universidad Nacional de Mar del Plata, presentando en el Informe de Tesis el trabajo desarrollado en esta línea de Investigación, bajo la dirección del Ph D. Nicolás Dámaso Patetta.

Referencias

[1] M. Berry & G. Linoff, G Data Mining Techniques: for marketing, sales, and customer relationship management (2a ed.) USA: Wiley Publishing, Inc, 2004

[2] P.S. Bradley & U.M. Fayyad Refining initial points for k-means clustering. In J. Shavlik, editor, Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98), San Francisco, CA, 1998

[3] C. Garcia Cambronero, I. Gomez Moreno Algoritmos de aprendizaje: KNN & Kmeans. Universidad Carlos III de Madrid, 2009. Recuperado de: www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf

[4] J. Hartigan & A. Wong A k-means clustering algorithm. Journal of the Royal Statistical Society, Series C (Applied StatisTIC), Vol. 28, No. 1, 1979. Recuperado de: http://www.jstor.org/stable/2346830.

[5] D. Huerta Muñoz Diseño de Planes eficientes para la segmentación de clientes con múltiples atributos. Tesis de Maestría de la Universidad Autónoma de Nuevo León. Facultad de Ingeniería Mecánica y Eléctrica. México, 2009.

[6] J.B. MacQueen Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, USA: University of California Press, 1967.

[7] L.R. Malbernat Innovación en educación universitaria: Factibilidad de incorporar actividades virtuales según las competencias docentes. Tesis de Maestría de Universidad Nacional de Mar del Plata. Facultad de Ciencias Económicas y Sociales. Argentina, 2012.

[8] UNESCO Las tecnologías de la información y la comunicación en la formación docente. Guía de planificación. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 2004 Recuperado de: http://unesdoc.unesco.org/images/0012/001295/129533s.pdf

[9] S. Vega Pons Combinación de resultados de Clasificadores no supervisados. Tesis de doctorado. Rep. Téc. Reconocimiento de Patrones. Serie Azul. Cuba: Centro de Aplicaciones de Tecnologías de Avanzada, 2011.

[10] E. Yolis, P. Britos, G. Perichisky & R. García-Martínez Algoritmos Genéticos Aplicados a la Categorización Automática de Documentos. Revista Electrónica de sistemas de Información. ISSN 1677-3071 Doi:10.5329/RESI, 2 (2), 2009. Recuperado de: http://revistas.facecla.com.br/index.php/reinfo/article/view/133/27.