Revista Ciencia en Desarrollo, Vol. 7 No. 1
ISSN 0121-7488 – Enero-Junio de 2016, pp. 15-20

Una técnica de clasificación con variables categóricas

A Classification Technique With Categorical Variables

J. A. Clavijo M.[a,*]
H. A. Granada D.[a]

Recepción: 15-oct-2015
Aceptación: 20-ene-2016

Resumen

Presenta un algoritmo de clasificación para elementos caracterizados por variables categóricas, usando *k-modas*, un algoritmo similar a *k-medias*. A la vez, se incluyen diagramas de flujo para la implementación del algoritmo en cualquier lenguaje de programación. También se presenta un ejemplo con datos reales que ilustra la propuesta.

***Palabras clave*:** conglomerado, distancia, disimilaridad, *k*-medias, *k*-modas.

Abstract

This paper presents a classification algorithm for elements characterized by categorical variables, using *k-modes*, a procedure analogous to *k-means*. At the same time we have included flowcharts for the algorithm implementation in any programming language. We also present a simple example, with real data, illustrating the proposal.

***Key words*:** Cluster, Distance, Dissimilarity, *k*-means, *k*-modes.

[a]Universidad del Tolima, Facultad de Ciencias, Departamento de Matemáticas y Estadística.
[*]Autor de correspondencia: jaclavijom@ut.edu.co

1. Introducción

La clasificación de individuos en diferentes conglomerados a partir de los valores que tome un conjunto de variables definidas sobre ellos es un procedimiento de gran interés en estadística, ya que tiene numerosas aplicaciones en las que se busca determinar segmentos muy homogéneos de una población.

El caso en que todas las variables que describen a los individuos sean de tipo numérico es ampliamente conocido [2, 4, 5], y se han proporcionado varias técnicas para formar conglomerados: unas de tipo jerárquico, como el *single linkage* o el *método de Ward*, y otras de tipo no jerárquico, como el *método k-means*. Sin embargo, el caso en que las variables observadas sean de tipo categórico ha sido menos estudiado y prácticamente no existen técnicas que de manera directa conduzcan a la formación de conglomerados. Podría citarse un método indirecto, subproducto del análisis de correspondencias, en el que se pueden calcular las coordenadas de los individuos y de las categorías sobre un *biplot* para agruparlos aplicando técnicas del caso numérico a dichas coordenadas.

El propósito de este artículo es presentar un método de clasificación que actúe directamente sobre los valores de las variables categóricas y agrupe los individuos basándose en la semejanza de los valores categóricos que ellos asumen. El método que se propone es una adaptación del método *k-means* de variables numéricas, utilizando el concepto de **moda**, en vez del concepto de media, idea que ha sido propuesta por varios autores, entre ellos [1, 6, 7], y que ha servido de inspiración para este trabajo.

2. Disimilaridad entre individuos

Todos los métodos de clasificación buscan reunir en un solo grupo los individuos que más se parecen entre sí de acuerdo con los valores que ellos asumen en las variables que se estudian. Si se consideran p variables $X_1, X_2, \ldots, X_p$, donde cada variable X_k tiene n_k categorías, cada individuo x_i se asocia con una p-upla $x_i = \left(c_{i_1 1}, c_{i_2 2}, \ldots, c_{i_p p}\right)$, donde $1 \leq i_k \leq n_k$ y $c_{ik} = c_{i_k k}$ es la categoría que dicho individuo adopta en la variable X_k con $k = 1, 2, \ldots, p$. Naturalmente, dos individuos se parecen más cuando coinciden en un número alto de categorías en las p-uplas correspondientes y se diferencian según el número de discrepancias que tengan.

Definición 1. Dados dos individuos $x_i = (c_{i1}, c_{i2}, \ldots, c_{ip})$ y $x_j = (c_{j1}, c_{j2}, \ldots, c_{jp})$ diremos que entre ellos hay una **discrepancia en la k-ésima variable** si $i \neq j$, la cual se representa mediante la métrica discreta:

$$\delta(c_{ik}, c_{jk}) = \begin{cases} 0 & si \quad i = j \\ 1 & si \quad i \neq j \end{cases}$$

Definición 2. Para todo par de individuos x_i, x_j definimos la **disimilaridad**:

$$d_{ij} = d(x_i, x_j) = \sum_{r=1}^{p} \delta(c_{ir}, c_{jr})$$

Observación 1. Dados dos individuos x_i y x_j, $d_{ij} = 0$ indica una *máxima semejanza* entre los dos individuos, mientras que $d_{ij} = p$ indica *máxima diferencia* o ningún parecido entre ellos.

3. Moda de un conjunto

Definición 3. Consideremos el conjunto $X = \{x_1, \ldots, x_n\}$ de n individuos descritos por p variables categóricas $X_1, \ldots, X_p$. Definimos una **moda** de X como una p-upla $Q = (q_1, \ldots, q_p)$ tal que $D(Q, X) = \sum_{i=1}^{n} d(x_i, Q)$ sea mínima [1].

Observación 2. Se garantiza la existencia de un mínimo para D, ya que es una suma finita de números enteros y está acotada por:

$$0 \leq D(Q, X) = \sum_{i=1}^{n} d(x_i, Q) \leq np$$

Ejemplo 1. Sea X el conjunto formado por seis individuos definidos por:

$$x_1 = (2,1,1) \quad x_2 = (2,3,3) \quad x_3 = (1,2,2)$$
$$x_4 = (1,1,1) \quad x_5 = (1,1,2) \quad x_6 = (2,2,3)$$

Estas triplas han sido definidas como respuesta a tres variables categóricas, así:

X_1 con dos categorías codificadas como 1 y 2.
X_2 con tres categorías codificadas como $1, 2$ y 3.
X_3 con tres categorías codificadas como $1, 2$ y 3.

Las posibilidades de respuesta son en total 18, y la moda de X será una terna Q que haga mínimo el valor $\sum_{i=1}^{6} d(x_i, Q)$. Como se puede ver en la tabla 1, hay 6 posibles modas para el conjunto dado, a saber: $(1,1,1)$, $(1,1,2)$, $(1,2,3)$, $(2,1,1)$, $(2,1,2)$ y $(2,1,3)$.

Tabla 1. Discrepancias d_{ij} para X y Y.

Q	x_1 211	x_2 233 y_1	x_3 122 y_2	x_4 111	x_5 112	x_6 223 y_3	$D(Q,X)$	$D(Q,Y)$
111	1	3	2	0	1	3	10	8
112	2	3	1	1	0	3	10	7
113	2	3	2	1	1	3	12	8
121	2	3	1	1	2	2	11	6
122	3	3	0	2	1	2	11	5
123	3	1	1	2	2	1	10	3
131	2	2	2	1	2	3	12	7
132	3	2	1	2	1	2	11	5
133	3	1	1	2	2	3	12	5
211	0	2	3	1	2	2	10	7
212	1	2	2	2	1	2	10	6
213	1	1	3	2	2	1	10	5
221	1	2	2	2	3	1	11	5
222	2	2	1	3	2	1	11	4
223	2	1	2	3	3	0	11	3
231	1	1	3	2	3	2	12	6
232	2	1	2	3	2	2	12	5
233	1	0	3	3	3	1	12	4

Es claro que el procedimiento descrito en el ejemplo anterior no es eficiente para encontrar una moda. Será particularmente engorroso cuando se manejen muchos individuos y un número grande de variables categóricas. Por ejemplo, una encuesta con solo 20 preguntas categóricas, cada una con 5 modalidades o categorías, exige un cálculo de más de 95 billones de sumas con tantos sumandos como individuos hayan contestado la encuesta, lo que genera un alto costo computacional!

Definición 4. Dado un subconjunto cualquiera $Y \subseteq X$ con $r \leq n$ elementos y una variable categórica X_j, para cada categoría c_{kj} de X_j se define **la frecuencia relativa** como el número:

$$fr\left(X_j = c_{kj}|Y\right) = \frac{n_{c_{kj}}}{r}$$

donde $n_{c_{kj}}$ es la cantidad de veces que c_{kj} pertenece a las p-uplas asociadas a los r elementos de Y.

Una manera práctica de obtener las frecuencias es mediante tablas de frecuencias relativas.

Ejemplo 2. Consideremos el conjunto $Y = \{y_1, y_2, y_3\} = \{x_2, x_3, x_6\}$ de $r = 3$ elementos (ver tabla 1), calculamos las frecuencias relativas:

$$fr(X_3 = c_{23}|Y) = fr(X_3 = 2|Y) = \frac{1}{3}$$

$$fr(X_2 = c_{12}|Y) = fr(X_2 = 1|Y) = 0$$

De manera similar se calculan las demás frecuencias relativas de las variables X_i en Y dadas por las tablas 2 y 3.

Tabla 2. Frecuencias Relativas de la Variable X_1 en Y.

X_1		
c_{i1}	$n_{c_{i1}}$	fr
$c_{11} = 1$	1	1/3
$c_{21} = 2$	2	2/3

Tabla 3. Frecuencias Relativas de las Variables X_2 y X_3 en Y.

X_2		
c_{i2}	$n_{c_{i2}}$	fr
$c_{12} = 1$	0	0
$c_{22} = 2$	2	2/3
$c_{32} = 3$	1	1/3

X_3		
c_{i3}	$n_{c_{i3}}$	fr
$c_{13} = 1$	0	0
$c_{23} = 2$	1	1/3
$c_{33} = 3$	2	2/3

El Teorema 1 referenciado en [1] proporciona una forma eficiente de encontrar una moda para los elementos de un subconjunto Y de X.

Teorema 1. El valor $D(Q,Y)$ es mínimo si y solo si $fr(X_j = q_j|Y) \leq fr\left(X_j = c_{kj}|Y\right)$ para $q_j \neq c_{kj}$ y $j = 1, \ldots, p$.

En esencia, este teorema nos dice que la moda Q para Y es la p-upla $Q = (q_1, \ldots, q_p)$, formada por las categorías $q_1, \ldots, q_p$ correspondientes a las máximas frecuencias en las tablas de frecuencia de cada variable.

Así, por ejemplo, para el conjunto Y del ejemplo anterior, Q debe ser la tripla $Q = (c_{21}, c_{22}, c_{33}) = (2,2,3)$, como se puede apreciar en la última columna de la tabla 1. Este vector Q es uno de los dos vectores que hacen mínima la suma $\sum_{i=1}^{3} d(y_i, Q)$ con $y_i \in Y$.

4. El algoritmo k-Modas

Ya que hemos establecido una metodología para encontrar modas en un conjunto de individuos determinados por variables categóricas, proponemos un algoritmo para agrupar dichos individuos en k conglomerados. El algoritmo que se propone es análogo al conocido método k-means para el caso de variables numéricas, pero usando modas en vez de medias. Es por esto que el mejor nombre que podemos elegir para identificar la propuesta es el de algoritmo de k-modas. Dicho algoritmo se define por medio de los cuatro pasos siguientes:

Paso 1: si se tienen n individuos en X que se pretenden agrupar en k clusters $(k < n)$, comenzamos por elegir k de dichos individuos para que actúen como núcleos de aglutinamiento. Hay varias formas de elegir estos núcleos: aleatoriamente, a partir de una lista predefinida o simplemente tomando los k primeros. Estos primeros núcleos actúan como modas transitorias. La última opción es la que adoptamos en este trabajo.

Paso 2: examinamos uno a uno los $n-k$ individuos restantes y asignamos cada uno al núcleo que le sea más cercano con la distancia d de la Definición 2. De esta manera vamos formando los diferentes grupos. En caso de empate en la distancia, el elemento en consideración se puede asignar aleatoriamente a uno de los núcleos que producen el empate. Una vez asignado un elemento a uno de los grupos, utilizamos el teorema visto anteriormente para actualizar la moda de dicho grupo. La nueva moda asumirá el papel de núcleo para el grupo en consideración.

Paso 3: una vez se hayan asignado todos los elementos de X a alguno de los grupos, se revisa la disimilaridad de cada elemento en cada grupo respecto a su propio núcleo y a los núcleos de los demás grupos. Si se encontrase un elemento que está más cercano al núcleo de un grupo que no es el suyo, reasignamos dicho elemento a ese grupo y actualizamos las modas tanto del grupo receptor como del emisor. Como antes, las nuevas modas actuarán como núcleos de estos dos grupos.

Paso 4: repetir el paso 3 cuantas veces sea necesario hasta que no se produzcan más cambios.

4.1. Implementación del método

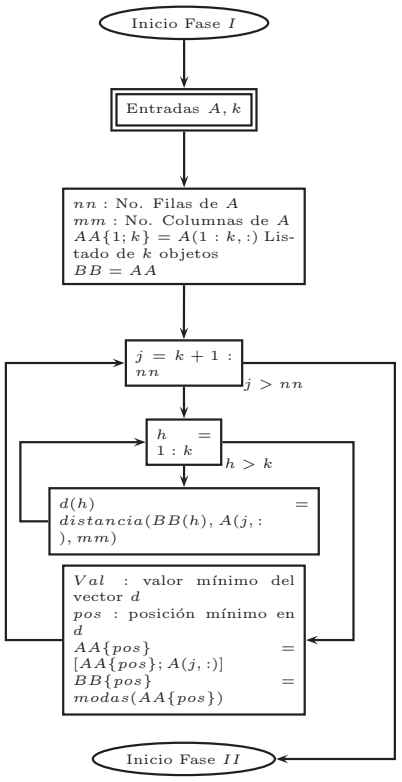
Como se puede apreciar, el algoritmo es relativamente simple, pero es demasiado dispendioso por la cantidad de veces en que se debe buscar una moda para actualizar una ya existente en cada conglomerado. Es este un proceso que difícilmente puede llevarse a cabo en forma manual, siendo necesario el uso de procedimientos computacionales. Se programaron en Matlab las rutinas necesarias para implementar el método anterior, que fue probado en la clasificación de 45 usuarios de los sistemas de salud del Tolima. La muestra fue recolectada en las ciudades de Ibagué y Espinal, mediante la aplicación de una encuesta de 7 preguntas con respuestas categóricas, relacionadas con la percepción que tienen los usuarios de las EPS que les prestan servicios médicos. Este trabajo fue realizado por las estudiantes Nataly J. Roa y Luisa Fernanda Pastrán, dentro de su trabajo de grado, relacionado con este tema [3]. En este trabajo presentaremos únicamente los diagramas de flujo de la programación del algoritmo, el cual se desarrolla en dos fases cuyos fines son los siguientes:

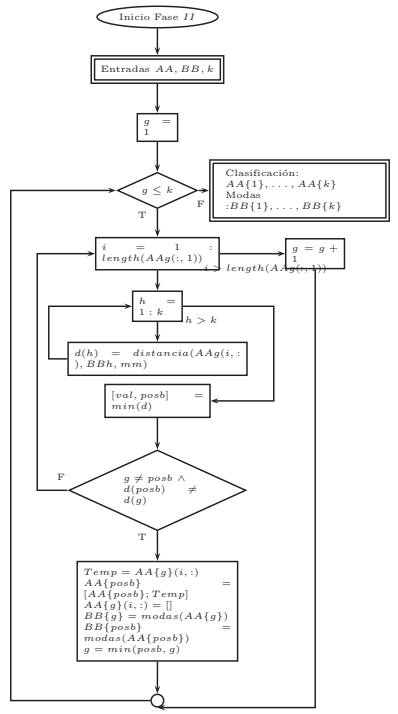
Fase 1. Búsqueda del núcleo más cercano para cada elemento y cálculo de la moda del grupo receptor.

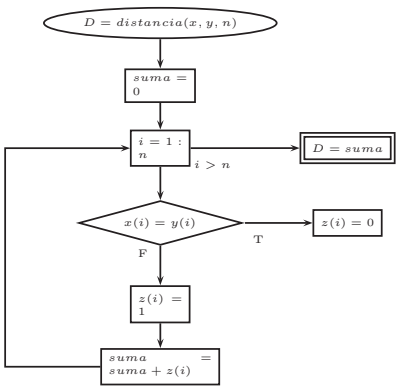
Fase 2. Comparación de todos los elementos de cada grupo con los diferentes núcleos para determinar si siguen en su grupo o si deben ser reasignados a otro grupo. Actualización de las modas tanto en el grupo emisor como en el grupo receptor en caso de alguna transferencia de elementos.

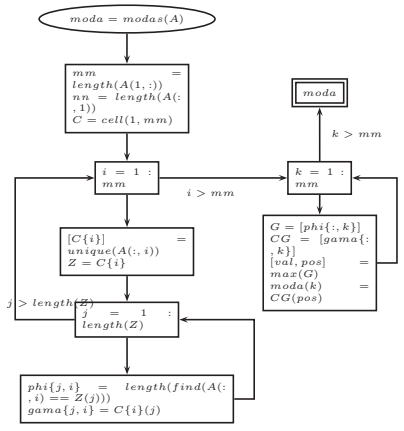
4.2. Diagramas de flujo

Como se mencionó anteriormente, el algoritmo de k-modas se realiza en dos fases, donde se hace necesario implementar una subrutina para búsqueda de modas y otra para el cálculo de disimilaridades. Los diagramas de flujo presentados en este trabajo fueron implementados en el software Matlab, y es por esta razón que damos por conocidas las funciones internas: cell, length, find, min, max y unique.









5. Aplicación del método k-Modas

La tabla 4 proporciona los resultados de aplicar una encuesta de 7 preguntas de tipo categórico a 45 usuarios de EPS de las ciudades de Ibagué y Espinal, en el Tolima [3].

Tabla 4. Encuesta d_{ij} para X y Y.

EPS	x_i	P_1	P_2	P_3	P_4	P_5	P_6	P_7
A_1	1	2	2	2	1	5	5	4
A_1	2	2	1	3	2	5	3	3
A_1	3	2	2	3	2	3	4	3
A_1	4	2	2	3	2	5	5	3
A_1	5	2	2	3	2	5	5	3
A_1	6	2	1	1	2	5	1	3
A_1	7	2	1	2	2	5	3	2
A_1	8	1	2	2	2	5	5	3
A_1	9	2	1	2	1	5	4	2
A_1	10	2	1	2	2	5	3	2
A_1	11	1	2	2	3	5	4	5
A_1	12	1	2	2	3	5	5	4
A_1	13	1	2	3	2	5	5	3
A_1	14	1	1	3	1	5	2	4
A_1	15	1	2	3	2	5	5	3
A_2	16	1	2	2	2	1	5	3
A_2	17	2	1	3	3	5	1	1
A_2	18	1	2	3	3	5	5	5
A_2	19	2	1	2	3	1	5	1
A_2	20	2	2	3	3	2	5	1
A_2	21	2	2	3	3	3	5	3
A_2	22	2	1	2	1	1	5	4
A_2	23	1	2	3	3	1	5	3
A_2	24	1	2	1	1	5	5	3
A_2	25	2	1	1	1	1	5	3
A_2	26	1	2	3	2	5	5	3
A_2	27	2	1	2	1	5	5	1
A_2	28	1	2	2	3	5	5	4
A_2	29	1	2	3	2	5	5	3
A_2	30	1	1	3	3	1	3	3
A_3	31	1	2	3	2	4	5	4
A_3	32	1	2	3	2	5	5	5
A_3	33	1	2	3	2	5	5	5
A_3	34	1	2	2	2	3	5	3
A_3	35	1	1	3	1	5	5	3
A_3	36	1	2	3	2	5	5	3
A_3	37	1	2	3	3	5	5	5
A_3	38	1	2	3	3	5	5	5
A_3	39	1	2	3	1	5	5	5
A_3	40	1	2	3	1	5	5	4
A_3	41	1	1	3	1	5	5	5
A_3	42	1	2	3	3	3	4	5
A_3	43	1	2	3	3	4	5	5
A_3	44	1	2	3	2	4	5	5
A_3	45	1	2	3	3	5	5	4

Los datos de la tabla 4 fueron procesados aplicando las rutinas antes mencionadas para realizar una clasificación en tres grupos, obteniéndose los resultados de la clasificación en las tablas 5, 6 y 7. En la tabla 8 se presentan las modas de cada uno de los grupos.

6. Conclusiones

Un examen de caracterización o tipología de estos grupos mediante la tabla de frecuencias de la tabla 9, muestra que el grupo mayoritario, con 33 individuos, está conformado por personas satisfechas con el servicio médico, que consideran que las instalaciones son buenas y que han conseguido sus citas médicas oportunamente en menos de una semana, que a la vez han recibido a tiempo sus medicamentos y que creen que la imagen de su EPS mejora cada vez más.

Tabla 5. Grupo 1: Usuarios Satisfechos.

P_1	P_2	P_3	P_4	P_5	P_6	P_7
1	2	2	2	5	5	3
1	2	2	3	5	4	5
1	2	2	3	5	5	4
1	2	3	2	5	5	3
1	1	3	1	5	2	4
1	2	3	2	5	5	3
1	2	2	2	1	5	3
1	2	3	3	5	5	5
2	2	3	3	3	5	3
1	2	3	3	1	5	3
1	2	1	1	5	5	3
1	2	3	2	5	5	3
1	2	2	3	5	5	4
1	2	3	2	5	5	3
1	2	3	2	4	5	4
1	2	3	2	5	5	5
1	2	3	2	5	5	5
1	2	2	2	3	5	3
1	1	3	1	5	5	3
1	2	3	2	5	5	3
1	2	3	3	5	5	5
1	2	3	3	5	5	5
1	2	3	1	5	5	5
1	2	3	1	5	5	4
1	1	3	1	5	5	5
1	2	3	3	3	4	5
1	2	3	3	4	5	5
1	2	3	2	4	5	5
1	2	3	3	5	5	4
2	2	3	2	5	5	3
2	2	3	2	5	5	3
1	1	3	3	1	3	3
2	1	3	2	5	3	3

Tabla 6. Grupo 2: Usuarios Insatisfechos.

P_1	P_2	P_3	P_4	P_5	P_6	P_7
2	1	1	2	5	1	3
2	1	2	2	5	3	2
2	1	2	2	5	3	2
2	1	3	3	5	1	1
2	1	1	1	1	5	3
2	1	2	1	5	4	2
2	1	2	3	1	5	1
2	1	2	1	1	5	4
2	2	2	1	5	5	4
2	1	2	1	5	5	1

Tabla 7. Grupo 3: Atípico.

P_1	P_2	P_3	P_4	P_5	P_6	P_7
2	2	3	2	3	4	3
2	2	3	3	2	5	1

Tabla 8. Modas de los 3 Grupos

Grupo	P_1	P_2	P_3	P_4	P_5	P_6	P_7
1	1	2	3	2	5	5	3
2	2	1	2	1	5	5	1
3	2	2	3	2	2	4	1

El grupo mediano, con 10 individuos, es un grupo de usuarios inconformes que han dado una baja calificación a las instalaciones y al sistema de citas, pues han tardado más de una semana para conseguirlas. En general, son individuos insatisfechos con el servicio médico. Finalmente, se formó un grupo atípico con dos individuos que no encajaron adecuadamente en ninguno de los dos grupos anteriores.

Tabla 9. Frecuencias por Grupo - Tipología.

Variable	Categoría	Satisfecho	Insatisfecho	Atípicos	Total
Comodidad, aspecto y funcionalidad del sitio de atención	1. Si	29	0	0	29
	2. No	4	10	2	16
Calidad del servicio de petición de citas	1. No	5	9	0	14
	2. Si	28	1	2	31
Dificultad para conseguir que lo atendieran	1. Muy difícil	1	2	0	3
	2. Un poco	6	7	2	15
	3. Fácil	26	1	0	27
Tiempo desde que solicitó el servicio y la fecha de atención	1. Mas de 6 dias	6	5	0	11
	2. De 4 a 6 dias	15	3	1	19
	3. De 1 a 3 dias	12	2	1	15
Grado de satisfacción con la entrega de los medicamentos	1. Muy insatisfecho	3	3	0	6
	2. Poco	0	0	1	1
	3. Indiferente	3	0	1	4
	4. Satisfecho	3	0	0	3
	5. Muy satisfecho	24	7	0	31
Calidad del servicio profesional (médico) recibido	1. Muy insatisfecho	0	2	0	2
	2. Poco	1	0	0	1
	3. Indiferente	2	2	0	4
	4. Satisfecho	2	1	1	4
	5. Muy satisfecho	28	5	1	34
Opinión sobre la evolución de su EPS (Camnio de imagen)	1. Pésima	0	3	1	4
	2. Mala	0	3	0	3
	3. Igual	16	2	1	19
	4. Mejor	6	2	0	8
	5. Excelente	11	0	0	11

Referencias

[1] Z. Huang, "Extensions to the $k-$Means Algorithm for Clustering Large Data Sets with Categorical Values Data Mining and Knowledge Discovery", *Kluwer Academic Publishers*, vol. 2, no. 3, pp. 283-304, 1998.

[2] A. C. Rencher, "Methods of multivariate analysis", *John Wiley & Sons*, vol. 492, 2003.

[3] N. Roa y L. F. Pastrán, "Una técnica de Clasificación con Variables Categóricas", *Trabajo de Grado*, *Universidad del Tolima*, Colombia, 2014.

[4] W. Dillon, and M. Goldstein, "Multivariate Analysis, Methods and Applications", *Jhon Wiley and Sons*, pp. 186-190, 1984.

[5] D. F. Morrison, "Multivariate Statistical Methods", *Mc Graw Hill*, pp. 389-391, 1990.

[6] B. Tian, C. A. Kulikowsky, G. Leiguang, Y. Bin, H. Lan, and Z. Chunguang, "A Global K-modes Algorithm for Clustering Categorical Data", *Chinesse Journal of Electronics*, vol. 21, no. 3, 2012.

[7] S. Mingoti, and R. Matos, "Clustering Algorithms for Categorical Data: A Monte Carlo Study", *International Journal of Statistics and Applications*, vol. 2, no. 4, pp. 24-32, 2012.