Segmentación del padrón de sujetos vulnerables en el Sistema de Asistencia Alimentaria

Juan Pablo Granados García[1], Rosario Baltazar[1], Silvia Quintana Vargas[2], Claudia Leticia Díaz Gonzáles[1], Martha Alicia Rocha Sánchez[1], Arturo Hernández Aguirre[3]

[1] División de Estudios de Posgrado e Investigación,
Instituto Tecnológico de León,
León, Guanajuato, México

[2] Hospital Regional de Alta Especialidad del Bajío,
León, Guanajuato, México

[3] Centro de Investigaciones en Matemáticas,
Departamento de Ciencias de la Computación,
Guanajuato, México

http://posgrado.itleon.edu.mx
artha@cimat.mx

Resumen. En este trabajo se llevó a cabo una caracterización de sujetos vulnerables en el Sistema de Asistencia Alimentaria del DIF del Estado de Guanajuato, esto aplicando algoritmos de agrupamiento. La segmentación se realizó con la información del Padrón 2013-2014. Se aplicó un proceso de limpieza y normalización de los datos. Además se creó un vector de características, al cual se le aplicaron algoritmos de agrupamiento como el Método de Ward y el K-means. Posteriormente se contrastó con algoritmos de clasificación de mínima distancia, KNN, Red Neuronal Artificial Back-propagation y Red Neuronal Artificial con PSO.

Palabras clave: minería de datos, segmentación, agrupamiento, clasificación.

1. Introducción

Los programas de ayuda alimentaria en México tienen como objetivo contribuir en la canasta básica de los beneficiarios para mejorar su nutrición. Los programas de alimentos y cocinas del Sistema Nacional para el Desarrollo Integral de la Familia (DIF) y de las Organizaciones No Gubernamentales (ONG) tienen una cobertura muy baja, por lo tanto, es importante hacer una distribución eficiente de los recursos para lograr beneficiar a los sujetos más vulnerables[23].

Una forma para lograr una distribución eficiente de los recursos es utilizar reconocimiento de patrones para generar grupos de beneficiarios con características similares. En el 2000 [25] se elabora un trabajo que resume y compara algunos de los métodos conocidos y utilizados en diversas etapas de un sistema de

reconocimiento de patrones. Además demuestra que el diseño de un sistema de reconocimiento requiere varias etapas: definición de clases, el medio de detección, la representación de patrones, la extracción de características y selección, el análisis y agrupamiento, el diseño del clasificador y el aprendizaje, la selección de muestras de entrenamiento y finalmente la evaluación del desempeño.

Posteriormente en el 2010 [27] se presenta un nuevo método para analizar la relación entre las probabilidades producidas por un modelo de clasificación y la variación de sus valores de entrada. El objetivo es aumentar la probabilidad de predicción de una clase dada. Mediante la exploración de los posibles valores de las variables de entrada.

Ahora bien la Escala Latinoamericana y Caribeña de Seguridad Alimentaria (ELCSA) presenta su validez interna y externa en 2010 [24]. Se recomienda su aplicación por la Secretaría de Seguridad Alimentaria y Nutricional (SESAN) y la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO). ELCSA responde a la necesidad de ampliar y mejorar la medición del hambre, usando métodos para medir directamente la experiencia en los hogares ante la inseguridad alimentaria y hambre.

En el 2013 [10] se realizó una investigación con respecto a la asistencia alimentaria en México que utiliza nuevas tecnologías que favorezcan o apoyen el análisis, resultados e implementaciones. Bajo el contexto anterior, se plantean modelos matemáticos que permitan el uso de algoritmos de clasificación y reconocimiento de los diferentes modelos de seguridad alimentaria.

En este trabajo de investigación se delimita en lo particular al Sistema de Asistencia Alimentaria del organismo DIF del Estado de Guanajuato. Una de las tareas del DIF es la identificación de la inseguridad alimentaria y hace hincapié en la reducción del hambre y la desnutrición en las familias con bajos recursos.

Con esta investigación se pretende generar grupos con características propias que nos permite conocer el comportamiento y la tendencia de los sujetos vulnerables, utilizando técnicas de minería de datos y reconocimiento de patrones. Por lo tanto, se convierten los datos en conocimiento útil. Se aporta nueva información de apoyo a la labor de identificación, para mejorar la distribución de los recursos y un control efectivo de los beneficiarios.

De esta manera es posible crear un perfil particular con las diferentes tendencias en base a los indicadores sociales de los sujetos y poder atender en su momento a ciertas áreas que requieran de un determinado apoyo más que otras.

2. Marco teórico

2.1. Agrupación de datos

Agrupación de datos o clasificación no supervisada es un proceso de asignación de un conjunto de registros en subconjuntos, llamados Clusters, de tal manera que los registros en el mismo grupo son similares y registros en diferentes grupos son muy diferentes [7,2].

2.2. Clasificadores no supervisados

Método de Ward: Es un algoritmo de agrupamiento jerárquico que divide un conjunto de datos en una secuencia de particiones anidadas representados por un diagrama de árbol o dendrograma [5,6]:

Paso 1: un nuevo cluster C_n está formado por la fusión de dos grupos en el conjunto inicial de clusters $F_0 = \{C_0, C_1, ..., C_{n-1}\}$. Luego, después de la etapa 1 y antes del paso 2, se tiene un conjunto de clusters $F_1 = \widetilde{F}_0 \cup \{C_n\}$ donde $\widetilde{F}_0$ es serie de clusters sin combinar en F_0. Si C_0 y C_1 tienen la distancia mínima entre todos los pares de grupos, entonces $C_n = C_0 \cup C_1$ y $\widetilde{F}_0 = \{C_2, C_3, ..., C_{n-1}\} = F_0 \setminus \{C_0, C_1\}$.

Paso 2: un nuevo grupo C_{n+1} está formado por la fusión de dos grupos en el conjunto de clusters F_1. Del mismo modo, dejamos $\widetilde{F}_1$ el conjunto de agrupaciones sin combinar en F_1. Luego después del paso 2 y antes del paso 3, tenemos un conjunto de clusters $F_2 = \widetilde{F}_1 \cup \{C_{n+1}\}$. El algoritmo continúa este proceso hasta que en el paso $n-1$ cuando los dos últimos grupos se fusionaron para formar el cluster C_{2n+2}. Después del paso $n-1$, tenemos $F_{n-1} = \{C_{2n+2}\}$, que contiene un solo cluster. El algoritmo se detiene después de la etapa $n-1$.

En el proceso anterior, se tiene $|F_0| = n$, $|F_0| = n$, $\quad |F_1| = n-1, ..., |F_{n-1}| = 1$, donde $|\quad .\quad |$ denota el número de elementos en el conjunto. Para decidir que grupos se fusionara, se tiene que calcular las distancias entre las agrupaciones. Para calcular la distancia entre un de cluster y un nuevo grupo formado por dos grupos Lance-Williams propone [30].

Antes del paso $i\,(1 \leq i < n-1)$, se tiene un conjunto de clusters F_{i-1}, que contiene $n-i+1$ clusters. Suponiendo que C_{i_1} y C_{i_2} tienen la menor distancia entre todos los pares de clusters en F_{i-1}. Entonces C_{i_1} y C_{i_2} se fusionarán para formar el cluster C_{n+i+1}. Para calcular la distancia entre un cluster $C \epsilon \widetilde{F}_{i-1} = F_{i-1} \setminus \{C_{i_1}, C_{i_2}\}$ es:

$$D\left(C_{i_1}, C_{i_2}\right) = \sqrt{\frac{2\left|C_{i_1}\right|\left|C_{i_2}\right|}{\left|C_{i_1}\right| + \left|C_{i_2}\right|}} d\left(c_{i_1}, c_{i_2}\right) \tag{1}$$

donde c_{i_1} y c_{i_2} representan los centroides de los clusters C_{i_1} y C_{i_2}, cuya distancia se está midiendo, $d\left(c_{i_1}, c_{i_2}\right)$ es una métrica que devuelve la distancia entre ambos centroides.

K-means: El algoritmo divide el conjunto de datos en k grupos $C_0, C_1, ..., C_{n-1}$, reduciendo la función objetivo 2 [5,2]:

$$J = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj} * \left[\left(\overline{x}_j - \overline{z}_k\right)^{\lambda}\right]^{1/\lambda} \tag{2}$$

Donde u_{kj} es igual a 1 si el j-ésimo punto pertenece al grupo k y 0 en caso contrario, $\overline{z}_k$ denota el centro del grupo k, $\overline{x}_j$ denota la j-ésimo punto de los datos. Los diferentes pasos del algoritmo K-means son:

Paso 1: Seleccionar K centroides de los grupos $z_1, z_2, ..., z_k$ de manera aleatoria para los n puntos $x_1, x_2, ..., x_n$.

Paso 2: Asignar el punto $x_1, i = 1, 2, ..., n$ al cluster $C_j, j\epsilon 1, 2, .., k$ si solo si $\left[(\overline{x}_i - \overline{z}_j)^\lambda\right]^{1/\lambda} < \left[(\overline{x}_i - \overline{z}_p)^\lambda\right]^{1/\lambda}$, $p = 1, 2, ..., K$, $j \neq p$. Los empates se resuelven de manera arbitraria.

Paso 3: Calcular nuevo centroides de los grupos $z_1^*, z_2^*, ..., z_k^*$ como sigue: $z_i^* = \frac{\Sigma_{x_j \epsilon C_i} x_j}{n_i}$, $i = 1, 2, ..., K$, donde n_i es el número de elementos que pertenecen al grupo C_j.

Paso 4: El proceso de partición se repite hasta que se cumpla una sentencia de paro: *a*) Los centros de los grupos no cambian, *b*) El valor J se hace menor que un umbral, *c*) El número máximo de iteraciones se han agotado. De lo contrario $z_i = z_i^*$, $i = 1, 2, ..., K$ y continuar desde el paso 2.

2.3. Reconocimiento de patrones

El reconocimiento de patrones es la disciplina científica cuyo objetivo es la clasificación de objetos en cierto número de categorías o clases. Dichos objetos pueden ser imágenes, señales o cualquier medida que pueda ser clasificada, los cuales se denominan patrones [8]. En el campo de reconocimiento de patrones se refiere a la detección automática de regularidades en los datos mediante el uso de algoritmos de computadora y con el uso de estas regularidades poder tomar acciones tales como la clasificación de los datos en diferentes categorías [9].

2.4. Clasificadores supervisados

Clasificadores de distancia mínima: Es el más utilizado para la clasificación de patrones de diferente tipo. Parte del hecho que las clases son linealmente separables. Se supone que la pertenencia de cada patrón se conoce y además se sabe el número de clases en las que el patrón puede ser clasificado [11,12]. Los vectores de características son asignados a clases de acuerdo a su ala distancia entre los puntos medios respectivos.

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \tag{3}$$

Clasificadores de tipo *k*-vecinos KNN: Considerando un conjunto de N puntos, $x_1, x_2, ..., x_N \epsilon R^l$ que provienen de una distribución estadística desconocida. El objetivo es estimar el valor de la función de densidad de probabilidad desconocida en un punto dado x. De acuerdo con la técnica de estimación vecino k-más cercano, se realizan los siguientes pasos [8]:

1. Elegir un valor de k.
2. Encuentra la distancia entre x y todos los puntos de entrenamiento x_i, $i = 1, 2, ..., N$.
3. Encontrar los puntos k-más cercanos a x.
4. Calcule el volumen $V(x)$ en el que los k-vecinos más cercanos.
5. Calcular la estimación por $p(x) \approx \frac{k}{NV(x)}$

Clasificadores Neuronales: Una red neuronal artificiales (RNA) es un procesador masivamente paralelo distribuido que es propenso por naturaleza a almacenar conocimiento experimental y hacerlo disponible para su uso [11]. Este mecanismo de conocimiento se parece al cerebro: *a*) Es adquirido a través del aprendizaje, *b*) Se almacena mediante la modificación del peso sináptico de las distintas uniones entre neuronas. Una RNA pueden ser usados para la clasificar patrones de clases lineal o no linealmente separables.

Back-propagation (BP) es un método de aprendizaje supervisado de redes neuronales que consta de una capa de neuronas de entrada, otra de neuronas de salida y opcionalmente una o varias capas de neuronas ocultas [29]. Tiene la forma de propagar hacia atrás el error, desde la capa de salida a la de entrada, modificando los pesos de las capas intermedias. Permitiendo así aprender mediante un conjunto de ejemplo y obteniendo una salida coherente para una entrada de salida.

Particle Swarm Optimization (PSO) es un método metaheurístico proporcional basado en poblaciones. Se basa libremente en el comportamiento de un grupo coordinado, como el de congregación de aves. Mantiene una colección de partículas virtuales donde cada partícula representa una potencial mejora de la solución a un problema, que en el caso de las RNA, es un conjunto de valores para los pesos y sesgos que reduzcan al mínimo el error entre los valores de salida calculados y valores de salida conocidos en un conjunto de datos de entrenamiento [28].

2.5. Indicadores sociales

Nivel Socioeconómico (NSE): Es la norma creada por la Asociación Mexicana de Inteligencia de Mercado y Opinión Pública (AMAI) [13,18], basada en análisis estadístico, que permite agrupar y clasificar a los hogares mexicanos en siete niveles, de acuerdo a su capacidad para satisfacer las necesidades de sus integrantes en términos de: vivienda, salud, energía, tecnología, prevención y desarrollo intelectual.

$$NSE = \sum_{i=1}^{10} V_i \tag{4}$$

Donde $V_{i=1}$ Televisiones a color, $V_{i=2}$ Computadoras, $V_{i=3}$ Número de focos, $V_{i=4}$ Número de autos, $V_{i=5}$ Estufa, $V_{i=6}$ Baños Completos, $V_{i=7}$ Regadera, $V_{i=8}$Tipo de piso, $V_{i=9}$Número de habitaciones, $V_{i=1}$ Educación del jefe de familia.

Nivel Vulnerabilidad (NV): Se define desde tres dimensiones críticas[14]: *a*) Como un efecto directo/resultado, *b*) Resultado de varios factores de riesgo, *c*) La incapacidad de manejar tales riesgos.

$$NV = \sum_{i=1}^{12} V_i \tag{5}$$

Donde $-V_{i=1}$ Contar discapacidad igual a 7, $-V_{i=2}$ Contar discapacidad igual a 6, $-V_{i=3}$ Contar discapacidad igual 999, $V_{i=4}$ Contar número de discapacidades, $V_{i=5}$ Enfermedad Grave, $V_{i=6}$ Mujer embarazada, $V_{i=7}$ Mujer lactante , $V_{i=8}$ Alcohol, $V_{i=9}$ Drogas, $V_{i=10}$ Menor de 5, $V_{i=11}$ Mayor de 65, $V_{i=12}$ Migrante.

Nivel Discapacidad (ND): Es la adjetivación de la deficiencia en el sujeto y con una repercusión directa en su capacidad de realizar actividades en los términos considerados normales para cualquier sujeto de sus características [15].

$$ND = -V1 - V2 \tag{6}$$

Donde $V1$ = Contar discapacidad igual 7, $V2$ = Contar discapacidad igual 6.

Nivel Seguridad Alimentaria (NSA): Se define como la garantía de que los individuos, las familias y la comunidad en su conjunto, accedan en todo momento a suficientes alimentos inocuos y nutritivos, principalmente producidos en el país en condiciones de competitividad, sostenibilidad y equidad, para que su consumo y utilización biológica les procure óptima nutrición, una vida sana y socialmente productiva, con respeto de la diversidad cultural y preferencias de los consumidores [16,18].

$$NSE = \sum_{i=1}^{15} V_i \tag{7}$$

Hacinamiento (HC): Número de integrantes de la familia entre número de cuartos, se considera hacinamiento valores superiores a 2 [18].

$$HC = \frac{Integrantes}{Cuartos} \tag{8}$$

Problemas de Salud (PS): La salud no es sólo la ausencia de afecciones o enfermedades, sino también es el estado de bienestar somático, psicológico y social del individuo y de la colectividad [21,20].

$$PS = V1 - V2 - V3 \tag{9}$$

Donde $V1$ = Contar el número de problemas de salud, $V2$ = Contar igual a 11, $V2$ = Contar igual a 10.

Más Vulnerable (MV): El DIF maneja la ecuación 10 para la obtención de un parámetro numérico que resuma al sujeto a ser seleccionado.

$$MV = \begin{cases} NSA + (388 - NSE) + HC + NV + ND + PS, m = 0 \\ NSA + (388 - NSE) + HC + NV + ND + PS - 28, m = 1 \end{cases} \tag{10}$$

Indice de Necesidades Básicas Insatisfechas (INBI): Bajo esta concepción de pobreza, la CEPAL (Comisión Económica para América Latina y el Caribe) diseñó el método de medición de las Necesidades Básicas Insatisfechas (NBI) [17], para clasificar los hogares como pobres y no pobres. Las personas que pertenecen a un hogar con una necesidad insatisfecha se consideran como pobres $INBI = 1$ y con $INBI > 1$ se califican en una situación de miseria o pobreza extrema.

$$INBI = VI + VS + HC + DE + IE \tag{11}$$

Donde VI = vivienda inadecuada que presenta piso de tierra, VS = vivienda sin servicios considerada sin agua por acueducto o sin conexión a alcantarillado o a pozo séptico, HC = hacinamiento, DE = dependencia económica de los hogares cuyo jefe tenga un nivel educativo inferior a tercero de primaria y tres o más personas por cada persona ocupada, IE = inasistencia escolar a los hogares en los cuales algún niño entre 7 y 11 años, pariente del jefe, no asista a algún establecimiento educativo.

Canasta Básica: Conjunto de alimentos cuyo valor sirve para construir la línea de bienestar mínimo [19]. Éstos se determinan de acuerdo con el patrón de consumo de un grupo de personas que satisfacen con ellos sus requerimientos de energía y nutrientes.

3. Resultados actuales

Para cada una de las técnicas y algoritmos utilizados se programaron en C# con C++: *a*) C# es la parte visual e interactiva que inicializa cada algoritmo y presenta los resultados generados, *b*) C++ llegan los parámetros necesarios para cada algoritmo, realiza cada uno de los cálculos y regresa los resultados.

Inicialmente, los datos usados para la realización de este estudio, correspondieron al Padrón 2013-2014 del DIF, que pertenecen a los posibles sujetos vulnerables registrados en el estado de Guanajuato. El Padrón 2013-2014 con tiene la información recolectada con el instrumento de identificación de sujetos vulnerables basado en ELCSA. Entonces se tienen un conjunto inicial de 8397 familias con 27255 registros de posibles sujetos vulnerables que conforman todas las familias.

Se hizo un preprocesamiento de los datos. En esta etapa, se realizó la limpieza, eliminando los datos fuera de rango, datos atípicos y aquellos considerados inconsistentes. Se encontraron 232 datos con una escala de seguridad alimentaria inconsistente que se salía de los estándares empleados por el ELCSA. Por tanto, se utilizó un método de estimación, el cual consiste en eliminar dichos registros, quedando entonces 27023.

Además la presencia de datos faltantes en diferentes registros, se optó el algoritmo Non-linear Interative Projections by Alternating Least-Squares (NIPALS) como método para estimar los datos faltantes. El algoritmo NIPALS obtiene

estimaciones de las componentes y de los vectores que permiten describir la matriz de datos y estimar los datos faltantes [3,4]. El algoritmo extrae una componente principal a la vez.

Se procedió a seleccionar las variables que mejor representen la variabilidad de los datos. Por lo tanto, se realizo un análisis con estadísticos descriptivos para la selección características con una varianza diferente de cero. Estas correspondieron al nivel escolaridad, tipo de ocupación, ingreso del jefe de familia, problemas de salud, tipo de discapacidad, tipo de vulnerabilidad, servicio médico, elementos de vivienda, violencia, integrantes en la familia, presencia de menores, gasto alimentario, seguridad alimentaria y canasta alimentaria.

Considerando que las variables pueden manejar diferentes escalas, que puede implicar a que aquellas con un mayor rango de valores les quiten importancia a otras con un menor rango, todas las variables consideradas fueron normalizadas. Se estandarizo los datos con respecto a la normalización Min-Max [1], es un enfoque simple donde se fijan los valores mínimos y máximos de las variables normalizadas entre $[0, 1]$.

Dada la gran dimensionalidad del problema por la cantidad de características, se hizo indispensable la reducción de estas, para posibilitar y hacer más eficiente el análisis. Se aplicó un Análisis de Componentes Principales (ACP) [4] con el cálculo de los valores propios de la matriz de correlaciones y generando un total de 25 componentes con una descripción de 89.334 %.

La medida de métrica de distancia seleccionada para la aplicación de los diferentes algoritmos de agrupamiento y clasificación fue la Minkowski de orden $\lambda = 3$. Esta es una medida general distancia [7] comúnmente utilizada.

3.1. Primer experimento

Se aplicó el algoritmo Método de Ward, se produjo el dendrograma de la figura 1. Seleccionar un dendrograma en un nivel particular produce una parti-

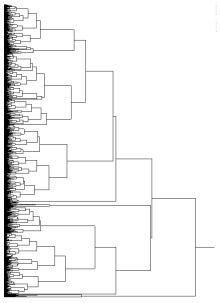

Fig. 1. Dendrograma generado por el método de Ward.

ción en grupos g disjuntos. Por lo tanto el dendrograma tiene un máximo 27023 grupos distintos. Para el análisis y busca del comportamiento de las diferentes

tendencias y definición de los grupos, se fue descendiendo por los niveles del dendrograma.

Se analizó cada uno de los grupos generados al descender por cada nivel, obteniendo de este modo las tendencias en los grupos mostrados en la tabla 2. Se generaran grupos explicativos del comportamiento y tendencias que ayudan a comprender el comportamiento de los datos.

Por otra parte, se utilizó el K-means para visualizar si persisten tales tendencias con los siguientes parámetros: grupos 18, iteraciones 100 y umbral 1. Se presentó una similitud ± 50, tabla 1 y se validó el modelo de comportamiento.

Tabla 1. 18 grupos y sus tamaños.

Cluster	Núm. Personas	Cluster	Núm. Personas
1	3704	**10**	379
2	3248	**11**	181
3	1577	**12**	132
4	3043	**13**	2442
5	1187	**14**	2174
6	40	**15**	2034
7	2353	**16**	1436
8	9	**17**	190
9	2776	**18**	118

Los grupos están en función de los diferentes indicadores sociales: violencia, vivienda inadecuada y sin servicios, dependencia económica, inasistencia escolar, INBI y canasta básica. Asi también con NSE, NV, ND, NSA, HC y PS usados por el DIF para obtener a los sujetos MV.

Cluster 1 Presencia de viviendas inadecuadas pero con servicios básicos y con hacinamiento. No existe la dependencia económica e inasistencia escolar. Tienen un INBI de pobre y pobre extremo, compuesta casi en su totalidad por NSE E, poca presencia NV, PS y nula por ND. Además la canasta básica con el gasto alimentario es muy parecidos.

Cluster 2 Presencia de familias grandes compuestas por 10 integrantes y con 2 menores en promedio por familia. Viven en viviendas inadecuadas, sin servicios básicos y con hacinamiento. Existe la dependencia económica, tiene un INBI como pobres extremos, compuesta casi en su totalidad por NSE E, presencia de NV por el número de menores, sin ND y algunos PS.

Cluster 3 Presencia de familias grandes compuestas por 9 integrantes y con 2 menores en promedio por familia. Viven en viviendas inadecuadas, sin servicios básicos y con hacinamiento. Existe la dependencia económica, tiene un INBI como pobres extremos, compuesta casi en su totalidad por NSE D, presencia de NV por el número de menores, sin ND y algunos PS.

Cluster 4 Presencia de viviendas inadecuadas sin servicios básicos y hacinamiento. No existe la dependencia económica e inasistencia escolar. Tienen

Tabla 2. Clusters generados.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Núm. Personas	3745	3298	1530	3044	1145	40	2335	9	2806	379	181	132	2440	2197	2098	1336	190	118
Inadecuada	3745	3298	1530	3032	1130	28	4	0	96	309	157	96	2432	2197	2083	1311	142	93
Sin Servicios	31	3292	1528	3040	1104	31	2087	0	72	273	138	101	1520	2197	64	0	114	85
Dependencia Económica	9	3262	1521	4	20	8	31	0	2746	133	50	29	40	0	0	7	44	26
Inasistencia Escolar	0	0	0	0	0	0	0	0	0	379	0	1	0	0	0	0	4	6
Pobre	1283	7	0	3	2	9	180	0	2794	0	16	15	655	0	2069	1309	24	26
Pobre Extremo	2462	3281	1530	3041	1143	31	2155	0	12	379	165	117	1785	2197	29	2	166	92
Núm. integrantes	5	10	9	5	4	4	4	3	5	5	4	4	3	2	2	3	5	4
Núm. menores	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	2	0	0
D+	0	0	0	22	0	0	37	1	129	0	0	0	0	0	0	476	0	0
D	243	182	1498	75	263	17	1905	0	611	96	36	25	568	0	231	852	55	64
E	3502	3116	32	2947	882	23	393	0	2066	283	145	107	1872	2197	1867	6	135	54
Si	3477	3232	1505	2297	867	18	68	0	580	291	90	48	2150	60	86	302	104	55
No	268	66	25	347	278	22	2267	9	2226	88	91	84	290	2137	2012	1034	86	63
0	3685	2146	927	3418	26	24	2120	3	1726	379	117	70	994	1069	1058	635	130	90
1	153	1068	603	6	1101	16	214	3	953	0	63	60	1444	1128	1040	697	60	28
2	7	14	3	0	18	0	1	0	127	0	1	2	164	0	0	4	0	0
0	3745	3207	1530	3041	1142	34	2335	4	2806	350	165	116	1	2197	2098	1324	155	93
1	0	91	0	3	2	3	0	0	0	21	15	14	2034	0	0	12	27	21
2	0	0	0	0	1	3	0	0	0	7	1	2	377	0	0	0	8	3
3	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	1	0	0	5	0	0	0	0	0
0	3567	3220	1338	2988	1117	23	2273	8	2684	349	145	103	1013	1427	1425	2	131	81
1	169	78	192	5	26	14	60	0	115	30	32	19	1124	769	670	802	48	26
2	9	0	4	3	2	3	2	0	7	0	4	10	269	1	3	473	9	10
3	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	53	2	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0
En Hogar	0	0	0	0	0	3	0	0	0	0	0	132	0	0	0	0	190	118
Que hizo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	115
Fallecimiento	0	0	0	0	0	40	0	0	0	0	172	132	0	0	0	0	0	20
Desaparición	0	0	0	20	0	0	0	0	0	0	152	132	0	0	0	0	0	18
Gasto Alimento	355	435	438	413	392	210	468	8500	355	388	439	340	261	228	225	300	396	328
Canasta Básica	355.63	357.77	359.03	354.36	354.97	358.39	594.4	319.25	357.93	351.86	357.82	355.48	362.37	371.72	367.22	343.93	348.99	334.60
Máximo	421	419	419	424	419	415	412	320	428	426	422.5	421.3	430	423	423	428	413	399.5
Mínimo	284	290	292	289.7	297	319	298.4	217.5	271	264.3	297	293	254.3	326	289.5	219.5	259.7	267.2
MV por DIF																		

un INBI de pobre extremo, está compuesta casi en su totalidad por NSE E y no presentan NV, ND y PS.

Cluster 5 Presencia de viviendas inadecuadas sin servicios básicos y hacinamiento. No existe la dependencia económica e inasistencia escolar. Tienen un INBI de pobre extremo, está compuesta por NSE D y E, presentan NV, pero no ND y PS.

Cluster 6 Presencia de sujetos registrados pero que están fallecidos en su totalidad.

Cluster 7 Presencia de viviendas adecuadas pero sin servicios básicos y sin hacinamiento. No existe la dependencia económica e inasistencia escolar. Tienen un INBI de pobre extremo, está compuesta casi en su totalidad por NSE D, poca presencia NV, ND y PS y la canasta básica es la más alta.

Cluster 8 Sujetos con NSE C y un gasto alimentario alto.

Cluster 9 Presencia de viviendas adecuadas, con servicios básicos y con poco hacinamiento. Existe dependencia económica, tiene un INBI como pobres, se componen por NSE D+, D y E. Presencia de NV=1,2 y la no existencia de ND y PS. La canasta básica con el gasto alimentario es muy parecidos.

Cluster 10 Presencia de inasistencia escolar en menores en su totalidad y considerados con INBI de pobre extremo.

Cluster 11 No existe la presencia de violencia en el hogar, pero existe fallecimiento familiar debido a un enfrentamiento armado o del crimen organizado y secuestro o víctima de desaparición forzada.

Cluster 12 Presencia de violencia en el hogar con la existencia de fallecimiento familiar y desaparición forzada del mismo en su totalidad.

Cluster 13 Presencia de viviendas inadecuadas sin servicios básicos y con hacinamiento, tienen un INBI de pobre y pobre extremo, está compuesta por NSE D y E. Gran presencia NV=1,2, ND=1, 2, 3, 4 y PS= 1, 2, 3.

Cluster 14 Presencia de viviendas inadecuadas sin servicios básicos y sin hacinamiento, tienen un INBI de pobre extremo, está compuesta en su totalidad por NSE E. Gran presencia NV=1 y PS=1.

Cluster 15 Presencia de viviendas inadecuadas con servicios básicos y sin hacinamiento, tienen un INBI de pobre, esta compuesta por NSE D y E. Gran presencia NV=1 y PS=1.

Cluster 16 Presencia de viviendas inadecuadas con servicios básicos y sin hacinamiento, tienen un INBI de pobre, está compuesta por NSE D+ y D. Gran presencia NV=1 y PS=1, 2, 3, 4.

Cluster 17 Presencia violencia en el hogar, sin fallecimiento y desaparición.

Cluster 18 Presencia violencia en el hogar, se realizó un tipo de acción para resolverlo, lo cual disminuye en el número de fallecimientos y desapariciones el hogar.

3.2. Segundo experimento

Para contrastar que existe tal comportamiento por parte de los datos en los nuevos grupos y que se pueda utilizar esta distribución generada para identificar la tendencia de un nuevo sujeto. Se implementaron los clasificadores Mínima

Distancia, KNN (K=1, 3, 5), una RNA entrenada con Back-Propagation y una RNA con PSO. Cada clasificador fue ejecutado 35 veces, en cada iteración se usó con un conjunto de datos de entrenamiento diferente. Se realizaron experimentos con un 20, 50 % de datos de entrenamiento de cada una de las clases.

Los Clasificadores Mínima Distancia y KNN (K= 1, 3, 5) utilizaron como métrica la distancia Minkowski de orden $\lambda = 3$. La estructura de la RNA entrenada con Back-propagation fue 25 neuronas en la capa de entrada, 49 en la capa intermedia, 18 en la capa de salida y el número de épocas igual a 10000. La RNA con PSO fue 25 neuronas en la capa de entrada, 100 partículas, 49 en la capa intermedia, 18 en la capa de salida y el número de épocas igual a 10000.

En la tabla 3 se muestran los porcentajes de clasificación mayor, medio y menor, obtenidos por cada clasificador durante las 35 ejecuciones de cada uno, en cada porcentaje de entrenamiento.

Tabla 3. Porcentajes de clasificación.

Clasificador	**% de entrenamiento**	**% Clasificación Mayor**	**% Clasificación Medio**	**% Clasificación Menor**
Distancia Minima	20.00	60.85	60.74	58.59
KNN = 1	20.00	95.82	95.38	94.90
KNN = 3	20.00	87.78	87.74	85.67
KNN = 5	20.00	61.22	60.48	60.17
RNA-BackPropagation	20.00	14.52	14.29	14.29
RNA-PSO	20.00	41.89	18.81	18.10
Distancia Minima	50.00	64.31	60.36	57.14
KNN = 1	50.00	96.80	96.71	95.38
KNN = 3	50.00	97.00	96.89	96.84
KNN =5	50.00	64.63	62.85	61.23
RNA-BackPropagation	50.00	14.29	14.29	14.29
RNA-PSO	50.00	41.91	34.91	28.58

Como se puede observar el clasificador KNN en sus 3 variantes obtuvo los mejores porcentajes de clasificación en las diferentes pruebas realizadas con 20 y 50 % de entrenamiento. Logrando un mayor porcentaje de clasificación del 97.00 % cuando K=3 en un 50 % de entrenamiento y un porcentaje menor de 60.17 % con K=5 en 20 % de entrenamiento.

La RNA-BackPropagation y RNA con PSO implementada obtuvieron los porcentajes de clasificación más bajos de todos los clasificadores. Teniendo además un costo computacional muy alto, ya que su proceso de entrenamiento y clasificación fue bastante largo.

4. Conclusiones

La elección de un vector de características para la aplicación de tecnificas de agrupamiento, influye directamente en los resultados del análisis, pues los

resultados dependen en gran medida de las variables que son consideradas.

Los algoritmos de clasificación, en lo particular el KNN pueden ayudar a identificar a los nuevos beneficiarios y colocarlos con su respectiva tendencia.

Los indicadores de vivienda inadecuada, sin servicios, dependencia económica e inasistencia escolar que corresponden al INBI con pobre y pobre extremo, describen mejor los resultados y muestran una idea más clara del comportamiento en los grupos. Estos indicadores hasta el momento no son usando por el DIF, pero están descritos por el CEPAL y la FAO.

El DIF para seleccionar un posible beneficiario utiliza al MV, esta ecuación no describe el comportamiento de la población. Ya que no se logró obtener separabilidad con respecto a esta.

Se obtuvo mayor información relevante, que puede ser empleada al momento de toma decisiones en el caso de si se otorgara algún tipo de apoyo y/o beneficio.

Los grupos generados dan una visión más clara del comportamiento de la población con respecto a sus formas de vida y por ello se pueden proponer diferentes tipos de acciones que se salen del Sistema de Asistencia Alimentaria del DIF. Este tipo de acciones pueden ser, por dar un ejemplo el mejoramiento de las necesidades básicas insatisfechas en el hogar o el apoyo a la inasistencia escolar.

Se presenta una canasta básica en los sujetos por debajo de lo establecido por Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). Ademas que no se consumen los alimentos mínimos necesarios.

Referencias

1. Samarasinghe, S.: Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. CRC Press., pp. 253–254 (2006)
2. Jain, A. K., Murty, M. N., and Flynn, P. J.: Data clustering: a review. ACM computing surveys (CSUR), 31(3):264—323 (1999)
3. Esbensen, K. H., Guyot, D., Westad, F., y Houmoller, L. P.: Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design. Multivariate Data Analysis, pp. 72–74 (2002)
4. Valencia, J. L., y Diaz-Llanos, F. J.: Regresión PLS en las ciencias experimentales. Madrid: Complutense, pp. 47–55 (2003)
5. Webb, A. R.: Statistical pattern recognition. Number 501-545. John Wiley y Sons (2011)
6. Rencher, A. C.: Methods of multivariate analysis. Vol. 492, John Wiley y Sons (2003)
7. Bandyopadhyay, S., Saha, S.: Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer Science y Business Media, 8–9, 68–72 (2012)
8. Theodoridis, S., Pikrakis, A., Koutroumbas, K., and Cavouras, D.: Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach. Number 1. Academic Press (2010)
9. Bishop, C. M. et al.: Pattern recognition and machine learning. volume 1, Springer New York (2006)

10. Cuevas-Ortuño, J. and Gómez-Padilla, A. Un modelo de asignaciónempaque de despensas personalizadas para bancos de alimentos: un sistema sujeto a condiciones nutricionales y logísticas. DYNA-Ingeniería e Industria, 88(5) (2013)
11. Rodríguez, Roberto; Sossa J. Humberto.: Procesamiento y análisis digital de imágenes. 1ª Edición, Alfaomega, México, pp. 19, 155-173 (2012)
12. Theodoridis, S., y Koutroumbas, K.: Pattern recognition, Fourth edition, pp. 30–31 (2009)
13. Asociación Mexicana de Inteligencia de Mercado y Opinión Pública *AMAI*, Consultado Marzo 2015 http://nse.amai.org/nseamai2/
14. FAO: La Seguridad Alimentaria: información para la toma de decisiones. Guía práctica. 1–4 (2011)
15. García, C. E., y Sánchez, A. S. Clasificaciones de la OMS sobre discapacidad. Boletín del RPD, 50, 15–30 (2001)
16. FAO: Ley marco derecho a la alimentación, seguridad y soberanía alimentaria. Aprobada en la XVIII Asamblea Ordinaria del Parlamento Latinoamericano, 16–23 (2012)
17. CEPAL: El uso de indicadores socioeconómicos en la formulación y evaluación de proyectos sociales. 15–32 (2001)
18. de la ELCSA, C. C.: Escala Latinoamericana y Caribena de seguridad alimentaria (ELCSA): manual de uso y aplicacione. 9–83 (2012)
19. Consejo Nacional de Evaluación de la Política de Desarrollo Social *CONEVAL*, Consultado Marzo 2015 http://www.coneval.gob.mx/Medicion/Paginas/Glosario.aspx (2015)
20. Organización Mundial de la Salud *OMS*, Consultado Marzo 2015 http://www.who.int/suggestions/faq/es/ (2015)
21. i Vives, S. S.: Vivir y morir en Alicante: higienistas e inversiones públicas en salud (1859-1923). Universidad de Alicante, 22 (2008)
22. Eberhart, R. C. and Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on micro machine and human science, volume 1, pp. 39–43, New York, NY (1995)
23. Morales-Ruán, M., Shamah-Levy, T., Mundo-Rosas, V., Cuevas-Nasu, L., Romero-Martínez, M., Villalpando, S., Rivera-Dommarco, J. A., et al.: Food assistance programs in mexico, coverage and targeting. salud pública de México, 55:S199–S205 (2013)
24. FAO: Validación de la Escala Latinoamericana y Caribeña para la Medición de la Seguridad Alimentaria (ELCSA) en Guatemala. Departamento de Nutrición Humana niversidad Estatal de Ohio. 1–18 (2010)
25. Jain, A. K., Duin, R. P. W., y Mao, J.: Statistical pattern recognition: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(1), 4–37 (2000)
26. Busygin, S., Prokopyev, O., y Pardalos, P. M.: Biclustering in data mining. Computers and Operations Research, 35(9), pp. 2964–2987 (2008)
27. Lemaire, V., Hue, C., and Bernier, O.: Correlation analysis in classifiers. Chapter From data mining to knowledge discovery: An overview, pp. 1–34 (2010)
28. Eberhart, R. C., y Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on micro machine and human science, Vol. 1, pp. 39–43 (1995)
29. Solé, R. V., & Manrubia, S. C.: Orden y caos en sistemas complejos (Vol. 2). Univ. Politèc. de Catalunya, 521 (2001)
30. Lemaire, V., Hue, C., and Bernier, O.: Correlation analysis in classifiers. Chapter From data mining to knowledge discovery: An overview. pp. 1–34 (2010)