BÚSQUEDA Y CARACTERIZACIÓN DE SUBGRUPOS DE POBREZA MEDIANTE LA APLICACIÓN DE ALGUNAS TÉCNICAS DE MINERÍA DE DATOS

Marta Sananes, [a] Elizabeth Torres, [a] Surendra P. Sinha [a] y Luis Nava Puente [b]

[a] Instituto de Estadística Aplicada y Computación, Universidad de Los Andes, Mérida, Venezuela
[b] Escuela de Estadística, Universidad de Los Andes, Mérida, Venezuela

Abstract

Data Mining or Knowledge Discovery in Databases comprises a variety of statistical and computational methods seeking to find relations and patterns of behavior in electronic repositories of data. Emergent relations and patterns can suggest the researcher causal explanations to be verified later or also they can suggest strategies of action to achieve some targeted changes. It is well known that poverty studies deal with a multifaceted problem, that has numerous causes, like economic conditions, environmental, cultural, political, demographics. Unable to face the complexity of the poverty phenomenon and because of its primary definition as scarcity of resources to satisfy basic needs, many studies tend to focus in the en economics aspects.

With this work we pretend, using computational intensive techniques that take in account huge amounts of data, to identify and characterize population subgroups that differ not just by its quantitative levels of poverty in economic terms but also by other associated characteristics, that we could, in general, all "ways" of being poor. Each found subgroup can suggest hypothesis on particular causes of its poverty and suggest, as well, recommendations about which strategies could be useful to improve or solve its condition. Coincidences in similar conditions may allow deduct strategies of general scope without loosing the point on group specificities.

For this study we used the Database of the Encuesta Nacional de Hogares sobre Medición de Nivel de Vida de Nicaragua (EMNV 2001)[2].

***Key Words**: Data Mininge, Knowledge Discovery Databases, Poverty, Pattern Recognition*

Resumen

La Minería de Datos, o Descubrimiento de Conocimiento en Bases de Datos, abarca una variedad de métodos estadísticos y computacionales para investigar la existencia de relaciones y patrones de comportamiento en almacenamientos electrónicos de datos. Relaciones y patrones emergentes pueden sugerir al investigador explicaciones causales que puedan ser verificadas posteriormente o bien pueden sugerir estrategias de acción para lograr ciertos objetivos de cambio. En el caso de los estudios de pobreza, se sabe que se trata de un problema multifacético resultante de numeras causas y condicionantes económicos, ambientales, culturales, políticos, demográficos. Ante la complejidad del problema de la pobreza y su definición primaria como carencia de recursos para satisfacer necesidades básicas, casi todos los estudios tienden a centrarse en los aspectos económicos.

Con este trabajo pretendemos, al emplear técnicas computacionales intensivas que toman en cuenta grandes cantidades de datos, lograr identificar y caracterizar subgrupos poblacionales que se diferencien no sólo por sus niveles cuantitativos de pobreza en términos económicos sino también por otras características asociadas, que podríamos en general denominar "maneras" de ser pobre. Cada subgrupo hallado puede sugerir hipótesis acerca de las causas particulares de su pobreza así como sugerir recomendaciones de cuáles estrategias pudieran ser útiles para mejorar o resolver su situación. Coincidencias en situaciones similares pudieran permitir deducir estrategias de alcance generalizado sin perder de vista las especificidades grupales.

Para el estudio utilizamos la base de datos de la Encuesta Nacional de Hogares sobre Medición de Nivel de Vida de Nicaragua (EMNV 2001))[2] .

***Palabras Claves**: Minería de Datos, Descubrimiento de conocimiento en Bases de Datos, Pobreza, Reconocimiento de patrones.*

Introducción

Tae-Wan Ryu´ y Eick [1] han resaltado las dificultades de aplicación de algoritmos comunes de *clustering* (conglomerados) a datos contenidos en Bases de Datos. Las aplicaciones existentes insumen conjuntos de datos almacenados en archivos "planos" mientras que las Bases de Datos almacenan los datos organizados en Tablas o Relaciones, conteniendo cada una un conjunto de *tuplas* (instancias) de valores de atributos (campos, variables), cada tupla unívocamente identificada por una clave de identificación constituida por uno o más atributos. Hay que tener presente que algunas operaciones de selección producen como resultado *bags* y no conjuntos.

Como señalan estos autores, las características apropiadas para considerar en la construcción de funciones de distancia requeridas por los algoritmos de clustering pueden estar diseminadas en diversas relaciones. Es un problema que en general afecta a todas las aplicaciones de Descubrimiento de Conocimiento en Datos (KDD), también llamado Minería de Datos (DM).

Por tanto, la primera fase de una investigación en DM es de preparación de datos, que incluye procesos de selección y construcción de los conjuntos de datos estructurados y almacenados de la forma requerida por la aplicación a utilizar. Tal como lo discute el artículo citado, es conveniente disponer de utensilios para la realización de esa primera fase, a partir de "vistas" extraídas de la BD.

Por su magnitud como encuesta de medición de niveles de vida así como por la disponibilidad de sus datos, se escogió para explorar la *Encuesta Nacional de Hogares sobre Medición de Niveles de Vida: Nicaragua 2001* (EMNV 2001) [2]. Los datos están registrados en varias versiones de la BD correspondientes a diferentes paquetes de software estadísticos (SAS, Stata) y en una versión plana en formatos coma-separado (CSV: comma delimited) con encabezado con identificación en secuencia de los atributos (variables, campos) registrados para las tuplas (instancias, observaciones). El equipo de desarrollo replicó en cada tabla información de control, lo que introduce mucha redundancia, en exceso de la mínima requerida por las claves de identificación y referenciales en el modelo relacional de datos subyacente.

Para la realización de este trabajo exploratorio se utilizó el paquete de software libre para DM WEKA [3], por su disponibilidad y calidad como producto académico. Sin embargo, resultó inadecuado para procesar el volumen de datos de la BD de EMNV 2001, ya que presenta algunas limitaciones, especialmente para lidiar con conjuntos voluminosos de datos, como lo reporta una revisión realizada en la School of Informatics, The University of Edinburgh [8]. Sólo aplicamos en este trabajo los algoritmos disponibles en WEKA para Clustering

WEKA requiere que los archivos de datos se presenten en un formato suyo particular, el formato ARFF (Attribute-Relation File Format). WEKA proporciona utensilios para preprocesado de datos, tales como "filtros" así como un editor que sirve para la primera fase preparatoria de datos siempre que ya estén en el formato requerido ARFF.

Para producir conjuntos de datos en formato ARFF a partir de las tablas o archivos originales en formato plano de EMNV 2001 se desarrollaron utensilios programados bajo el modelo Orientado a Objetos, incluyendo también utensilios para la realización de procesos previos, como juntar o mezclar (*join, merge*) tablas, seleccionar atributos, (*project*), recodificar variables, seleccionar (*select*) instancias (tuplas, líneas) según condiciones valores especificados de ciertas variables y para convertir al formato ARFF. Se aspira completar progresivamente un paquete generalizado (***DMUtiles***) y hacerlo libremente disponible a la comunidad de usuarios en general y en particular a la de WEKA.

La sección 1 contiene una descripción resumida de la BD EMNV 2001. La sección 2 contiene una descripción breve de WEKA y de los procesos de Clustering a utilizar. La sección 3 contiene la descripción de los utensilios de preparación de datos desarrollados. La sección 4 reseña las tablas elegidas de la BD de EMNV 2001 para trabajar en esta primera exploración usando WEKA. La sección 5 reseña los resultados de los algoritmos aplicados. La sección 6 presenta las conclusiones y perspectivas de continuación

1. Descripción de la Base de Datos de EMNV 2001

El portal del proyecto *Living Standards Measurement Study (LSMS)* [9] mantenido por el *Development Economics Research Group (DECRG) of the World Bank* (Banco Mundial) da acceso a toda la información de la Encuesta EMNV 2001 de Nicaragua, incluyendo las distintas versiones de la Base de Datos, formato del formulario de entrevista y extensa documentación.

En la documentación la página **==== INDICE ====.htm** contiene una tabla descriptiva de todos los documentos disponibles y enlaces a ellos.

El documento **mandb.pdf** (MANUAL DEL USUARIO DE LA BASE DE DATOS) describe la organización de la BD, dando los nombres de las tablas y una breve descripción del contenido en referencia a las secciones del formulario de encuestas.

El documento **Nicaragua 2001 Codebook.pdf** describe en extenso la definición de cada tabla o archivo de datos, con nombres de campos, longitud, tablas de códigos para campos de variables categóricas o nominales.

El documento **bolhogar.pdf** contiene el formulario de la encuesta.

Tanto la documentación como la BD se pueden descargar de la página de EMNV 2001, después de completar y remitir el 'Data Use Agreement Form'. En el sitio se puede consultar información adicional.

2. Descripción de WEKA

"Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License." Así se presenta WEKA en su portal web. (http://www.cs.waikato.ac.nz/ml/weka/)

WEKA y el libro "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations" de Witten y Frank [4] son producciones complementarias de la Universidad de Waikato, Nueva Zelandia.

WEKA implementa los siguientes algoritmos de Clustering:

Cobweb: *"Algoritmo para clustering incremental, agregando una instancia por vez. Produce un árbol en el cual las hojas representan instancias, el nodo raíz representa el conjunto completo de instancias y las ramas los clusters y subcluster hallados. Puede haber hasta un máximo de tantos subclusters como instancias en el conjunto de datos".*[13]

DensityBasedClusterer: *"Para este algoritmo, los cluster se consideran como regiones en el espacio de los datos con alta densidad separados por regiones de baja densidad. Las regiones pueden tener cualquier forma y cualquier distribución interna de los puntos. Para cada instancia de prueba de la calsificación se calcula una estimación de pertenencia a cada cluster en forma de una distribución de probabilidad. En algunos casos lo más interesante puede ser detectar outliers."*[14]

EM (Expectation-Maximization) *"El algoritmo EM asigna a cada instancia una distribución de probabilidad de pertenencia a cada cluster. El algoritmo puede decidir cuantos clusters crear basado en validación cruzada o se le puede especificar a priori cuantos debe generar. Utiliza el modelo Gaussiano finito de mezclas, asumiendo que todas los atributos son variables aleatorias independientes."* [15]

FarthestFirst

"Implements the "Farthest First Traversal Algorithm" by Hochbaum and Shmoys 1985: A best possible heuristic for the k-center problem"

SimpleKMeans

"K-means [10] es uno de los algoritmo de aprendizaje no supervisado más simples para aplicar al problema de *clustering*. A priori se asumen k clusters cada uno con su "*centroide*". Se busca que queden lo más lejos posibles unos de otros. A continuación, cada instancia del conjunto de datos es asociado al *centroide* más cercano formándose así el *clustering* inicial. Se recalcula la ubicación de los *centroides* como centros de masa de los clusters formados. El proceso continúa hasta que no haya más cambios. El algoritmo trata de minimizar una función objetivo de errores cuadráticos." [17]

3. Descripción de Utensilios Programados

El paquete de utensilios de preparación de datos se diseñó y programó bajo el modelo Orientado a Objetos. El diseño de clases tuvo como elemento de partida la identificación como Clase del concepto "Tabla de Códigos" para datos cualitativos (nominales o categóricos) o, equivalentemente, como relación distinguible en el modelo conceptual de datos relacional. Este concepto ya lo habíamos usado en trabajos anteriores [5, 6]. Cada tabla de códigos es identificada por un nombre único y almacena pares de valores (código, categoría). Admite posibilidad de recodificación si varios códigos se asocian a una misma

categoría. La idea es construir para cada BD a minar, si no lo tiene, un repositorio de Tablas de Códigos.

El siguiente concepto considerado es el de Variables, que se representa como clase que almacena una lista de variables que puede representar todas las variables o campos de una tabla de datos escogida para participar en el proceso de minado o bien ser una selección de ellas (*proyección*) o bien ser el subconjunto de los campos que conforma la clave primaria en la tabla. Almacena la lista de nombres de Variables con un *Tipo* asociado y longitud de campo. El Tipo es ´**NUMERIC**´ si la variable es de tipo cuantitativa o la referencia por nombre a la Tabla de Códigos que la codifica, si es de tipo cualitativo. Dado que el paquete actual está dirigido a la construcción de archivos en formato ARFF no se incluyó más detalle para el tipo cuantitativo pero una versión más general (***DMUtiles***) deberá en ese caso incluir también el concepto de "Rango" ya usado también en los trabajos citados.

El Gráfico 1 muestra el diseño de las clases básicas para el desarrollo del paquete de preparación de datos (***DMUtiles***), cuya primera versión se preparó para aplicación en este trabajo.

El diagrama de clases fue construido utilizando el utensilio “ModelMaker” [12] versión 6 incluido en la versión Demo de Delphi 7 (Delphi7 Trial) de Borland ® [11].

Con estas clases como fundamentales se diseñaron las demás clases que conforman el paquete de aplicaciones utilitarias en su primera versión (***DMUtiles 0***):

1 ***SelectTXT***: Construcción a partir de tabla original de datos en formato TXT con encabezado de nombres de variables (columnas, atributos, campos), de estilo MS Excel ®, previa proyección o selección de variables y selección de líneas o tuplas según condiciones, de archivo en formato TXT con encabezado de variables manteniendo el estilo MS Excel ®. Las condiciones pueden ser de las formas *Excluir si / Incluir solo si*. En caso de colisión de condiciones Excluir/Incluir, prevalece la inclusión.
2 ***MergeTXT***: Mezcla (join, merge) tablas de datos originales con selección previa de variables (proyección)y selección de líneas o tuplas según condiciones, para construcción de archivo en TXT con encabezado de variables (Estilo MS Excel ®)
3 ***ConstruARFF***: Construcción a partir de tabla original de datos, previa proyección o selección de variables y selección de líneas o tuplas según condiciones, de archivo en formato ARFF.
4 ***MergeTXT***: Mezcla (join, merge) de tablas de datos originales con selección previa de variables, para construcción de archivo en formato RAF
5 ***ConvertARFF***: Conversión simple a formato ARFF con proyección y recodificación

Otras operaciones de preparación o preproceso de datos pueden realizarse dentro del programa **Explorer** de WEKA.

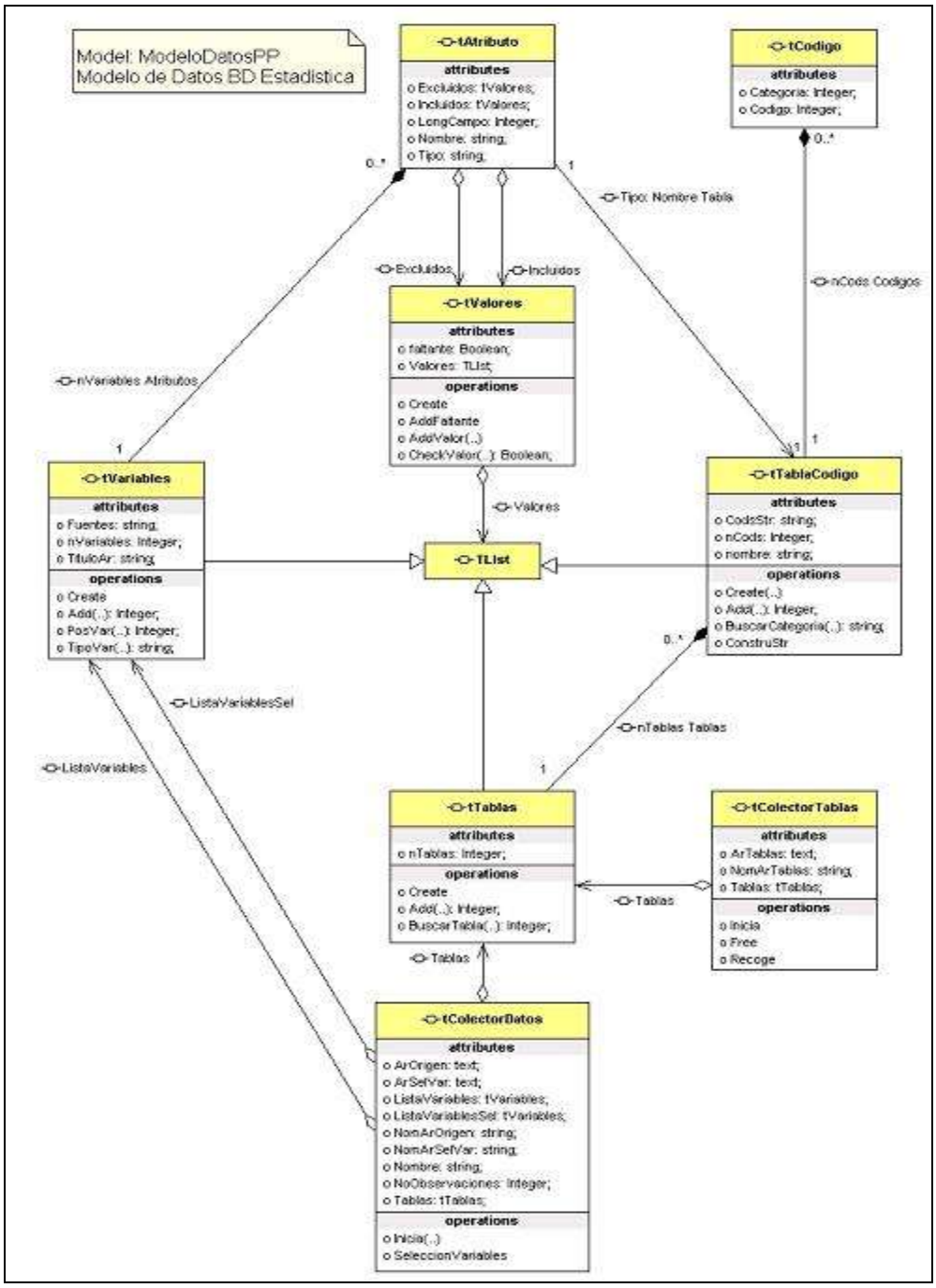


Gráfico 1. Diagrama de Clases básica DMUtiles

4. Selección de tablas de la BD de EMNV 2001

En las BD de EMNV 2001 todas las variables tanto cuantitativas como cualitativas tienen valores numéricos. En el caso de las cualitativas el *Codebook* documenta el significado de cada valor cualitativo.

Para facilitar la comprensión de las salidas de las aplicaciones de los algoritmos de WEKA conviene que las categorías de los datos nominales o cualitativos, sean valores textuales y no números. Por eso, usando el concepto de Tabla de Códigos, fue posible recodificar todas las variables cualitativas. Sin embargo, si bien la conversión la realiza la aplicación automáticamente así como la generación de las declaraciones en formato ARFF, se requiere realizar un trabajo previo de construcción del archivo con las Tablas de Código y con las selecciones de variables para cada tabla original escogida para participar en el proceso de minado. La magnitud de este trabajo dependerá de cada diseño de BD original así como de la documentación disponible.

En el caso de EMNV 2001 el trabajo se realizó a partir de la documentación. Tanto las Tablas de Códigos como los archivos de selección de variables se construyeron a partir del *Codebook* [7], documento en formato PDF, con el utensilio de "cortar" texto del Adobe Reader.

Una vez "pegados" los contenidos deseados -descripciones de archivos, variables y codificación- manualmente, también cortando y pegando, se construyeron los archivos "TablaCodigos.txt" y los archivos de selección de variables para cada tabla escogida.

Ejemplo de documentación de variables copiada del Codebook:

```
Variable Information:
Name Position
I00A Numero de formulario 1
Measurement level: Unknown
Format: F4 Column Width: Unknown Alignment: Right
I00B Numero del hogar 2
Measurement level: Ordinal
Format: F2 Column Width: Unknown Alignment: Right
Value Label
1 Hogar principal
2 Segundo hogar
3 Tercer hogar
4 Cuarto hogar
5 Quinto hogar

..............
S8P17A Pago dinero/bienes el trabajo realizado por no miembro hogar
45
Measurement level: Ordinal
Format: F1 Column Width: Unknown Alignment: Right
Value Label
1 Si
2 No
9 Ignorado
```

Ejemplo de tabla de códigos colocada en archivo "TablaCodigos.txt":

```
:Tabla 0
TipoHogar
1 HogarP
2 Hogar2
3 Hogar3
4 Hogar4
5 Hogar5
:Fin Tabla

:Tabla 1
UrbanoRural
1 Urbano
2 Rural
:Fin Tabla

:Tabla 2
SiNo
1 Si
2 No
9 ?
:Fin Tabla
```

Nótese como en la tabla se indica recodificar el código 9 de la Tabla SiNo (Ignorado en la codificación original) como el carácter ´?´, que es el símbolo de WEKA para datos faltantes.

Para construir un archivo de selección de variables se copia el nombre y longitud de cada variable a seleccionar del *Codebook* y se asigna el tipo, NUMERIC definido en WEKA para variables cuantitativas o el nombre de la tabla de códigos correspondiente si es variable cualitativa. En ambos archivos las líneas que comienzan con ´:´ son interpretados en la clase ColectorTablas como líneas de comando.

Ejemplo de variables seleccionadas del archivo original EMNV06:

```
:INFO: EMNV06 NEGOCIO PARTE A.sav
:SOURCE: Nicaragua EMNV 2001 - Banco Mundial
:Variable Information:

I00A Numero de formulario 1
Tipo: NUMERIC 4

I00B Numero del hogar 2
Tipo: TipoHogar 2
:Incluidos
1
2
:Fin Incluidos

I05 Area de residencia 10
Tipo: UrbanoRural 1
```

```
S8P7A Todo el negocio/actividad es de los miembros del hogar 22
Tipo: SiNo 1

S8P8 Razon para iniciar el Negocio/actividad 24
Tipo: RazonNegocio 2
:Excluidos
?
7
:Fin Excluidos

S8P9 El Negocio/actividad era: 25
Tipo: LocalNegocio 1

S8P10 Cuantos meses funciono/trabajo Negocio/actividad 26
Tipo: NUMERIC 2
:Excluidos
?
:Fin Excluidos
```

Con la definición de tablas en el archivo "***TablaCodigos.txt***" y de variables seleccionadas en el archivo de selección correspondiente, el utensilio ConstruARFF construye el archivo en formato ARFF, incorporando comentarios, las declaraciones WEKA ´**@relation**´, ´**@attribute**´, ´**@data**´ y genera las línea de datos con los valores de las variables seleccionadas y recodificadas.

Ejemplo de un archivo ARFF generado con parte del archivo original **EMNV06**:

```
% Title:  EMNV06 NEGOCIO PARTE A.sav
%
% Sources:  Nicaragua EMNV 2001 - Banco Mundial
%

@relation MisDatos

@attribute I00A NUMERIC
@attribute                                                              I00B
{Hogar_principal,Segundo_hogar,Tercer_hogar,Cuarto_hogar,Quinto_hogar}
@attribute S8P7A {Si,No}
.........................................

@DATA

1,HogarP,Si,Independencia,Local,12,?,1,126,No,1,0,750,Ampliar,trabajadore
s,Comercializacion,Si,1,BancoPrivado,Efectivo,20000,26000,Meses,42000,No,
?,No,?,No,?,Si,30000,Mucha_competencia,?,?,MuchaCompetencia,Managua
2,Segundo_hogar,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
,?,?,?,?,?,Managua
3,HogarP,Si,NoConsiguio,ViviendaSin,5,Si,1,48,No,0,0,?,SinCambios,Ninguno
,?,No,?,?,?,?,?,?,1200,No,?,Si,150,No,?,Si,700,VentasBajas,MuchaCompetenc
ia,?,VentasBajas,Managua
.........................
```

Sección 5. Aplicación de algoritmos y análisis de resultados

Para este primer estudio exploratorio de la BD de la Encuesta EMNV 2001, se preparó con los utensilios descritos de ***DMUtiles***, versión 0, un archivo ARFF juntando datos de los archivos originales **EMNV01** (sección 1 de la encuesta y parte de la 8), EMNV06 (segunda parte de la sección 8 que contiene información recopilada sobre negocios de miembros de las familias) y de **CONSING** (que contiene información resumida de la Encuesta, incluyendo la clasificación según los límites de ingreso considerados de las familias como *Pobre extremo, Pobre no extremo y No Pobre*. Se excluyeron los *No Pobres*. La tabla resultante (archivo ARFF) contiene 4614 observaciones y después de proyección, unas 80 variables.

El paquete WEKA no pudo procesar ese archivo. Otras pruebas con archivos de tamaño similar generaban igualmente mensajes de error detectando falsamente valores nominales no definidos. Para poder aplicar los algoritmos se tuvo que reducir tanto el números de variables como el número de observaciones. Se redujo a 21 el número de variables y a 263 el número de observaciones, lo cual es evidentemente muy insuficiente.

Con esta tabla se probaron los algoritmos *EM, SimpleKMeans y FarthestFirst.* Los dos últimos generan clusters representados por *centroides*, conformados por un individuo con ciertos valores de las variables consideradas.

Por ejemplo, el Cluster 3 utilizando el algoritmo SimpleKMeans comprende el 82% de las observaciones y su *centroide* presenta los siguientes valores:

```
Cluster 3
        Mean/Mode:  Urbano Zinc Accesible   1.7439 ConEscritura TuberiaFuera
        Std Devs:   N/A      N/A      N/A   1.0281 N/A           N/A

        FueraVivienda Queman Electrica NoTiene   1.125  FaltaCredito NoConsiguio
          N/A          N/A      N/A      N/A        0      N/A       N/A

        SeDesplaza BancoPrivado No 155.6195 VentasBajas PobreNoExtremo
           N/A      N/A         N/A  58.6086 N/A             N/A
```

(Zona Urbana, Techos de Zinc, Vía de Acceso accesible, Número promedio de cuartos disponibles en la vivienda, Vivienda propia con escrituras, Toma de agua fuera de la vivienda, Sanitarios fuera de la vivienda, Quema la basura, Tiene Luz eléctrica, No tiene teléfono, Tiempo medio a la escuela más cercana en horas, Razón para cerrar último negocio: falta de crédito, Razón para iniciar el negocio: no consiguió trabajo asalariado, Forma de hacer el negocio: se desplaza por las calles, obtuvo crédito de Banco Privado, No consume en el hogar los bienes producidos, Promedio valor bienes consumidos, Problema que afectó más al negocio: ventas bajas, es Pobre no extremo)

Suponemos que una caracterización semejante de un grupo basada en datos suficientes

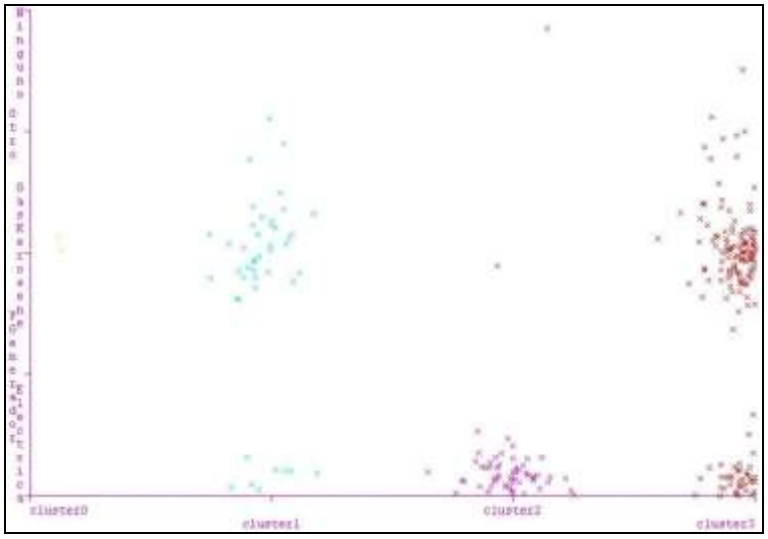
puede guiar un diseño de políticas públicas o de acciones privadas.
Otro ejemplo proviene de la aplicación del algoritmo EM. En este caso el cluster 3 con 60% de las observaciones se puede describir en base a las distribuciones calculadas para las distintas variables en cada grupo, de las cuales se muestra un pedazo:

Cluster: 3 Prior probability: 0.5866

Attribute: I05
Discrete Estimator. Counts = 17.51 142.71 (Total = 160.21)
Attribute: S1P7
Discrete Estimator. Counts = 102.36 34.94 6.08 11.94 7.78 1.11 (Total = 164.21)
Attribute: S1P9
Discrete Estimator. Counts = 80.34 57.83 23.05 1 (Total = 162.21)
Attribute: S1P13
Normal Distribution. Mean = 1.4331 StdDev = 0.5869
Attribute: S1P16
Discrete Estimator. Counts = 66.9 61.16 1.8 1.98 14.84 10.01 8.52 1 (Total = 166.21)
Attribute: S1P20
Discrete Estimator. Counts = 3.97 34.23 20.72 47.23 49.27 1.01 8.79 1 (Total = 166.21)
Attribute: S1P32
Discrete Estimator. Counts = 1.87 158.35 (Total = 160.21)
Attribute: S1P35
Discrete Estimator. Counts = 1.88 60.65 10.9 86.88 1.68 2.22 (Total = 164.21)

Se puede interpretar que para cada atributo los valores estimados mayores proporcionan el conjunto de las características representativas del grupo.

WEKA también produce gráficos ilustrativos de los resultados. El siguiente muestra como se distribuyen las observaciones en los cluster formados respecto a la variable S1P38 (Con que tipo de alumbrado cuenta principalmente este hogar), notándose el predominio en el grupo 3 de uso de *GasKeroseneCandil.*



Sección 6. Conclusiones

- Consideramos que, con software suficientemente poderoso para realizar minería de datos en la Encuesta EMNV 2001 o en otros estudios de alcance similar, se pueden obtener caracterizaciones significativas de grupos poblacionales, especialmente de los ubicados en situación de pobreza, con el fin de identificar las necesidades más resaltantes y poder concertar políticas públicas o mixtas con sectores privados y con las propias comunidades afectadas.

- WEKA es un excelente paquete para docencia de Minería de Datos o Machine Learnig como sus autores prefieran describirlo. Para uso extensivo resulta insuficiente.

- Es necesario probar la utilización de otros paquetes o utensilios para DM y en particular para *clustering*.

- Hay un campo de desarrollo de software en esta área de DM. Entre las posibilidades están: continuar el desarrollo del paquete auxiliar para preparación de datos (***DMUtiles***), investigar en el área de algoritmos de *clustering*, implementar nuevos algoritmos o realizar implementaciones poderosas de algoritmos publicados, por ejemplo, de los algoritmos programados en WEKA, integrar algoritmos en paquete de DM.

Bibliografía

[1] Tae-Wan Ryu´ y Christoph F. Eick. A database clustering methodology and tool, Information Sciences Volume 171, Issues 1-3 , 4 March 2005, Pages 29-59

[2] INSTITUTO NACIONAL DE ESTADISTICAS Y CENSOS INEC PROYECTO MECOVI - NICARAGUA.
http://www.worldbank.org/html/prdph/lsms/country/ni2001/ni01home.html

[3] Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand. (http://www.cs.waikato.ac.nz/ml/weka/)

[4] Ian H. Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann. San Francisco, 1999, 2000

[5] Sananes, Marta y Elizabeth Torres. Un Ambiente para Análisis de Datos. Economía, Revista Anual de la Facultad de Ciencias Económicas y Sociales de la Universidad de Los Andes. Mérida, 1996.

[6] Entralgo, E., Prototipo Procesador de Encuestas PPE, Licenciatura en Estadística, IEAC, ULA, 1989.

[7] Documentos de EMNV 2001, LSMS, Banco Mundial
http://www.worldbank.org/html/prdph/lsms/country/ni2001/ni01home.html

[8] Software for the data mining course. School of Informatics, The University of Edinburgh. http://www.inf.ed.ac.uk/teaching/courses/dme/html/software2.html

[9] LSMS home: http://www.worldbank.org/html/prdph/lsms/index.htm ->
LSMS Documents, Questionnaires, and Data Sets:
http://www.worldbank.org/html/prdph/lsms/docs.htm -> enlace data sets:

http://www.worldbank.org/html/prdph/lsms/guide/select.html -> buscar en la tabla Nicaragua 2001, columna DATA (enlace on web):
http://www.worldbank.org/html/prdph/lsms/country/ni2001/ni01home.html -> Data Agreement Form

[10] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

[11] Borland ® http://www.borland.com/delphi/

[12] ModelMaker Tools VB. The Netherlands. Borland Technology Partner.
http://www.modelmakertools.com/

[13] http://grb.mnsu.edu/grbts/doc/manual/COBWEB.html

[14] http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/
Prof. Dr. Hans-Peter Kriegel, Ordinarius Ludwig-Maximilians-Universitaet Muenchen.

[15] http://grb.mnsu.edu/grbts/doc/manual/Expectation_Maximization_EM.html

[16] Mathematics of Operations Research, 10(2):180-184, citado por Sanjoy Dasgupta. http://www.cs.ucsd.edu/~dasgupta/papers/hier-talk.ppt)

[17] http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html
Matteo Matteucci. PhD Student at Politecnico di Milano.
Department of Electronics and Information