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

Marta Sananes
Surendra P. Sinha
Elizabeth Torres
Luis Nava Puente
Instituto de Estadística Aplicada y Computación
Escuela de Estadística
Universidad de Los Andes
Mérida, Venezuela

Mayo, 2005

Objetivos

- Buscar relaciones y patrones emergentes que pueden sugerir explicaciones causales a ser verificadas posteriormente o bien que pueden sugerir estrategias de acción para lograr ciertos objetivos de cambio.

- Identificar y caracterizar sub grupos 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.

Recursos

- Base de datos de la Encuesta Nacional de Hogares sobre Medición de Nivel de Vida de Nicaragua (EMNV 2001).

  INSTITUTO NACIONAL DE ESTADISTICAS Y CENSOS INEC PROYECTO MECOVI - NICARAGUA.

http://www.worldbank.org/html/prdph/lsms/country/ni2001/ni01home.html

- Se utilizó el paquete de software para Minería de Datos: ***WEKA***

  Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand.

http://www.cs.waikato.ac.nz/ml/weka/

Fases

- Preparación de datos

  Desarrollo de utensilios para proyección o selección de variables, selección de observaciones, recodificación y conversión de datos (***DMUtiles***)

- Selección de software/algoritmos

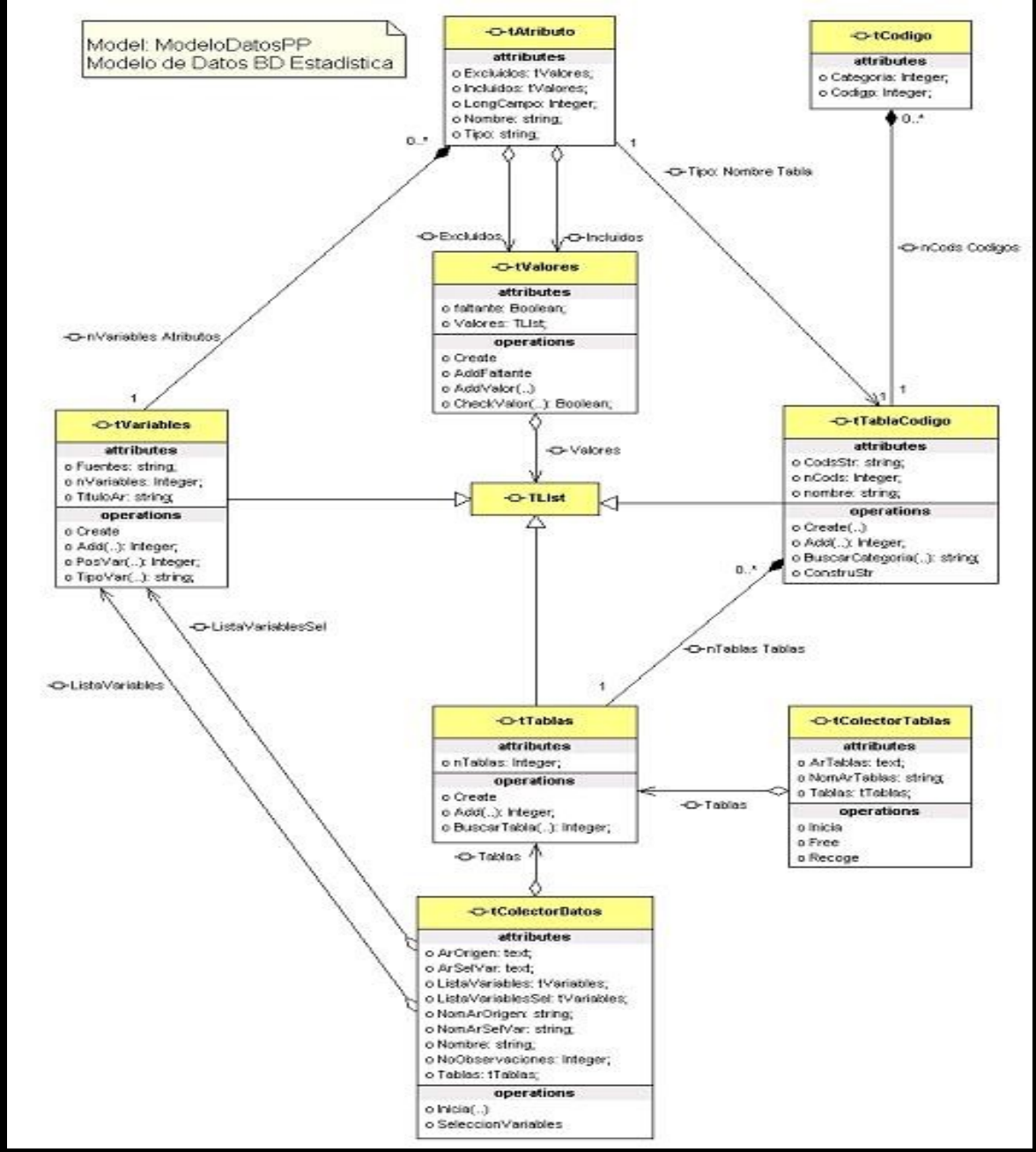
- Aplicación de algoritmos

- Discusión de resultados

Paquete de preparación DMUtiles, *versión 0*

- ***SelectTXT***: Proyección o selección de variables y selección de líneas o tuplas según condiciones, de archivo en formato TXT con encabezado de variables e el estilo MS Excel ®. Las condiciones pueden ser de las formas *Excluir si / Incluir solo si*.
- ***MergeTXT***: Mezcla (join, merge) tablas de datos con proyección y selección de líneas o tuplas según condiciones, para construcción de archivo en TXT con encabezado de variables (Estilo MS Excel ®)
- ***ConstruARFF***: Construcción a partir de tabla de datos, previa proyección y selección de líneas o tuplas según condiciones, de archivo en formato ARFF de WEKA.
- ***MergeTXT***: Mezcla (join, merge) de tablas de datos originales con selección previa de variables, para construcción de archivo en formato AARF
- ***ConvertARFF***: Conversión simple a formato ARFF con proyección y recodificación

Clases básicas para DMUtiles



Hecho con ModelMaker, versión 6 incluido en (Delphi7 Trial) de Borland ®

Descripción de la Base de Datos de EMNV 2001

- Proyecto *Living Standards Measurement Study (LSMS)* 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.
- 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.
- 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'.

Ejemplo de documentación de variables (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 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
```

Ejemplo de variables seleccionadas de archivo EMNV06

```
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
```

Otras variables seleccionadas del archivo EMNV06:

```
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
```

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."

(http://www.cs.waikato.ac.nz/ml/weka/)

Algoritmos para Clustering en WEKA

- ***Cobweb***
- ***Density Based Clusterer***
- ***EM*** *(Expectation-Maximization)*
- ***Farthest First***
- ***Simple K Means***

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".

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

Density Based Clusterer

"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 clasificación 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 extremos (outliers)."

http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/

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."

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

Farthest First, Simple K Means

FarthestFirst

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

http://www.cs.ucsd.edu/~dasgupta/papers/hier-talk.ppt

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."

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

Ejemplo de archivo ARFF para WEKA

```
% 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,MuchaCompet
   encia,?,VentasBajas,Managua
.........................
```

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
```

Valores descriptivos del centroide

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

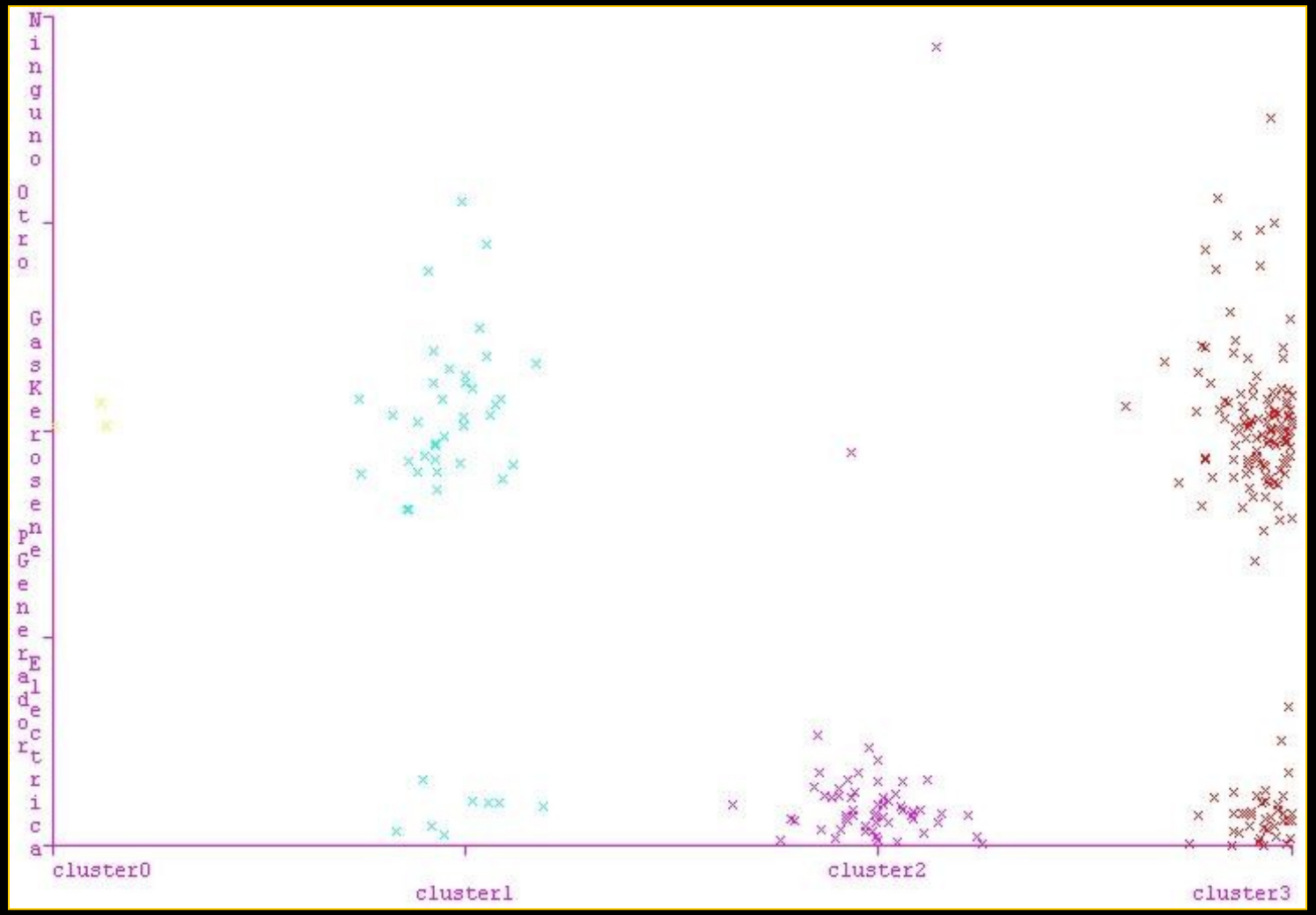
Es Pobre no extremo

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)
```

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.



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.

Conclusiones...

- 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.