Análisis Simbólico de Datos:una potente herramienta para Big Data

Adriana Mallea[1],Myriam Herrera[2], and María Inés Lund[2]

[1] *Departamento de Matemática, Universidad Nacional de San Juan, San Juan, Argentina*
[2] *Instituto de Informática, Universidad Nacional de San Juan, San Juan, Argentina*

E-mail: lamallea@ffha.unsj.edu,ar; {mherrera, mlund}@iinfo.unsj.edu.ar

Abstract Los datos simbólicos, introducidos por Edwin Diday en los ochenta, se ocupan del análisis de datos con variabilidad intrínseca que debería ser tenida en cuenta. En minería de datos, análisis multivariado de datos y estadística clásica los elementos analizados generalmente son entidades individuales, para las cuales se graba un valor individual de cada variable. Por ejemplo, individuos descriptos por edad, salario, nivel educativo, etc. Pero cuando los elementos de interés son clases o grupos de algún tipo, como los ciudadanos que viven en una ciudad determinada, modelos de autos en lugar de vehículos específicos, etc.; hay variabilidad inherente en los datos. Reducir esta variabilidad mediante medidas de tendencia central, tales como media aritmética, mediana o moda, lleva obviamente a una pérdida de información importante.

El análisis de datos simbólicos proporciona un marco que permite representar datos con variabilidad, usando nuevos tipos de variables. Los datos simbólicos se pueden representar usando los arreglos usuales en forma de matrices, pero en los cuales los elementos de cada celda no son valores numéricos reales individuales, sino conjuntos finitos de valores, intervalos o, de forma más general, distribuciones.

En los últimos años surgió el término **Big Data**, refiriéndose a conjuntos de datos tan grandes y complejos que se vuelven difíciles de procesar, en un tiempo razonable, con aplicaciones tradicionales de análisis de datos. El análisis simbólico de datos, al ofrecer la posibilidad de agregación de datos al nivel de granularidad elegido por el usuario mientras se mantiene la información sobre la variabilidad intrínseca, desempeña un papel importante en este contexto.

En el presente trabajo se desarrollan algunos conceptos fundamentales de la teoría de objetos simbólicos y se emplean metodologías del análisis de tales objetos en la Encuesta Permanente de Hogares, correspondiente al tercer trimestre del año 2016. En particular se trabaja con los hogares del Gran San Juan, región de Cuyo con dos propósitos: caracterizar a éstos hogares y comparar a los encuestados en cuanto a su estado social, en relación a variables de interés, en particular su Nivel de Estudios. Para el primer objetivo se define como objeto simbólico el hogar; mientras que para el segundo propósito se trabaja con objetos de tipo eventos. Se destaca la posibilidad de trabajar con bases individuales relacionales a fin de responder a diferentes propósitos de estudio, como también la generación de objetos simbólicos que son descriptos por variables simples, multivaluadas, probabilísticas y del tipo intervalo.

Con este trabajo se logra comprobar que los Datos Simbólicos son una herramienta de gran utilidad para el manejo y análisis de grandes volúmenes de datos, por lo que consideramos una herramienta fundamental para dos grandes áreas de la Ciencia de datos: Data Mining y Machine Learning.

Keywords Objeto simbólico, Variables Simbólicas, Data Mining.

1 Introducción

El Análisis de Datos Simbólicos permite la extensión de la Estadística a la Estadística de las intenciones o conceptos y más concretamente la extensión de problemas, métodos y algoritmos de análisis de datos clásicos a datos simbólicos. Según Diday el Análisis de Datos Simbólicos crea un puente entre la Estadística y

el Aprendizaje Automático.

Diday [1,2] formaliza los conceptos de intención y extensión, debidos a Arnauld y Nicole (Arnauld y Nicole, 1662). La intención de un concepto constituye su descripción, mientras que la extensión es el conjunto de individuos cuya descripción es acorde a la del concepto. La intención, que se representa por un objeto simbólico, se describe por los datos simbólicos y por un mecanismo de reconocimiento de los individuos de la extensión.

Diday introduce los objetos simbólicos y presenta una formalización que permite tratar conocimientos más ricos que los datos habituales y establece una relación con el modelo clásico de Análisis de Datos [3]. Un objeto simbólico representa una intención, un concepto y se define, en términos generales, como una conjunción de valores, o conjuntos de valores que pueden ser ponderados. Constituye una descripción en intención de una clase de individuos que constituyen la extensión.

Los objetos simbólicos representan conceptos, entendidos como la intención y extensión del mismo. La intención de un concepto representa las propiedades que lo definen y que lo hacen distinto de los demás conceptos. La extensión de un concepto se compone de los individuos que se definen por el concepto o que cumplen las propiedades que definen el concepto. Se describen por variables y datos simbólicos y proporcionan un mecanismo de vuelta a bases de datos o conjuntos de individuos en el sentido de conocer aquéllos que se adecuan o relacionan con las descripciones simbólicas representadas por los objetos (las intenciones), según determinadas relaciones que también forman parte de las intenciones.

El Análisis de Datos Simbólicos es una generalización de las técnicas de Análisis de Datos aplicadas a matrices de datos simbólicos. Las definiciones y notación de variables simbólicas y objetos simbólicos han estado en constante evolución desde sus inicios [1, 2, 3, 4, 5].

2 Datos Simbólicos

2.1 Preliminares

Sea $\Omega = \{\omega_1, ..., \omega_k\}$ un conjunto de individuos, sea $\mathcal{Y}$ un conjunto o dominio de posibles valores observados y $E = \{e_1, ..., e_n\}$ un conjunto de objetos. Como casos particulares más frecuentes se tiene que E es un subconjunto de Ω o un subconjunto de las clases de Ω, es decir, $E \subseteq \Omega$ o $E \subseteq P(\Omega)$. En el segundo caso, los datos simbólicos correspondientes, describen clases de individuos de Ω.

La descripción de un elemento $e \in E$ por una variable con dominio $\mathcal{Y}$ puede darse por: un elemento del conjunto $\mathcal{Y}$ (variable clásica), un subconjunto de elementos del conjunto $\mathcal{Y}$, un intervalo de valores del conjunto $\mathcal{Y}$ donde se ha definido un orden, un subconjunto de elementos del conjunto $\mathcal{Y}$ donde cada uno de ellos es ponderado por un peso o modo.

La matriz de datos en el Análisis Datos Simbólicos es la matriz $[X]$, cuyas filas representan n unidades u objetos del conjunto E descriptos por un vector de variables simbólicas $X = (X_1, ..., X_p)$. Es decir, las celdas de una fila se corresponden con los datos simbólicos descritos por el vector X aplicado a un elemento de E.

A continuación, se definen los distintos tipos de variables y datos simbólicos.

Variable conjunto valuada

Definition 1. *Se dice que X es una variable conjunto valuada si es una aplicación:*

$$X : E \to P(\mathcal{Y})$$
$$e \to X(e)$$

- $X(e)$ es la descripción de un elemento $e \in E$ en $P(\mathcal{Y})$ dada por la variable X

- $P(\mathcal{Y})$ es el conjunto de descripciones de los elementos de E.

En caso que $|X(e)| = 1$ (donde $|A|$ denota la cardinalidad del conjunto A) para todo $e \in E$ se denomina monovaluada. Se llama multivaluada se $1 < |X(e)| < \infty$ para todo $e \in E$. Se puede tratar de una variable categórica o cuantitativa.

Se puede extender la definición anterior al caso multivariante.

Variables modales probabilistas

Sea $\mathcal{Y} = \{z_1, ..., z_x\}$ y sea $\mathcal{M}(\mathcal{Y}) =$

$\{q : q$es una distribución de probabilidad definida en $\mathcal{Y}\}$,

el conjunto de descripciones modales probabilistas de elementos de E. Una descripción $q \in \mathcal{M}(\mathcal{Y})$ se define como:

$$q : \mathcal{Y} \to [0,1]$$
$$z_i \to q(z_i) \quad \text{con} \sum_{i=1,...,x} q(z_i) = 1$$

Se identifica el **dato simbólico** o descripción simbólica q con

$$q \equiv (z_1 q(z_1), ..., z_x q(z_x))$$

Definition 2. *Se dice que X es una variable modal probabilista definida en E, si es una aplicación*

$$X : E \to \mathcal{M}(\mathcal{Y})$$
$$e \to X(e) = q_e$$

tal que dado $e \in E$ le asocia $X(e) = q_e$ donde q_e es una distribución de probabilidad en el conjunto $\mathcal{Y}$ de posibles valores de observación, completado por una $\sigma - álgebra$

$X(e)$ es la descripción modal probabilista (en $\mathcal{M}(\mathcal{Y})$) del elemento $e \in E$ dada por la variable modal probabilista X.

En el caso en que $E \subseteq \mathcal{P}(\Omega)$, la variable X es un descriptor modal probabilista de clases de individuos de Ω y $\mathcal{M}(\mathcal{Y})$ el conjunto de las descripciones modales probabiliastas de clases de Ω, o de los elementos de $\mathcal{P}(\Omega)$.

La definición de variable modal probabilista se puede extender a una variable modal cuyos modos asociados a las categorías de $\mathcal{Y}$ son frecuencias o pesos. Así también al caso multivariante.

2.2 Objetos Simbólicos

Sea el conjunto $E = \{e_1, ..., e_n\}$ de elementos descriptos por p variables simbólicas $X_1, ..., X_p$ definidas en E con dominios finitos $\mathcal{Y}_1, ..., \mathcal{Y}_p$. En general la variable simbólica X_j es una aplicación $X_j : E \to \mathcal{D}$, siendo $\mathcal{D}$ un conjunto de descripciones de elementos de E asociado al conjunto o dominio $\mathcal{Y}$. El conjunto $\mathcal{D}$ puede ser: $\mathcal{Y}$ en el caso de que X_j sea una variable monoevaluada, $\mathcal{P}(\mathcal{Y})$ si X_j es una variable simbólica multievaluada o $\mathcal{M}(\mathcal{Y})$ si X_j es modal. Se dice que una descripción $d \in \mathcal{D}$ está asociada al conjunto o dominio $\mathcal{Y}$.
Se definen relaciones de dominio entre las descripciones. En particular para un par de descripciones se tiene:

Definition 3. *Sean $\mathcal{D}$ y $\mathcal{D}'$ dos conjuntos de descripciones de clase asociados a un mismo dominio, $\mathcal{D} \times \mathcal{D}'$ su producto cartesiano, una relación de dominio $\mathcal{R}$ definida en $\mathcal{D} \times \mathcal{D}'$ es una aplicación:*

$$\mathcal{R} : \mathcal{D} \times \mathcal{D}' \to \mathcal{L}$$
$$(d, d') \to \mathcal{R}(d, d') := [d\mathcal{R}d']$$

que a cada par de descripciones $(d, d') \in \mathcal{D} \times \mathcal{D}'$ le asocia un valor, denotado por $[d\mathcal{R}d']$ que mide el grado de adecuación o conexión de ambas descripciones. $\mathcal{L}$ es el conjunto de comparación de descripciones. El valor $[d\mathcal{R}d']$ es el nivel de relación entre las descripciones d y d'.

La relación de dominio es booleana si el conjunto $\mathcal{L} = \{0, 1\}$. Si $\mathcal{L} = [0, 1]$, se dice que $\mathcal{R}$ es una relación difusa.

Definition 4. *Sea una colección de relaciones de dominio $(\mathcal{R}_1, ..., \mathcal{R}_p)$, cada $\mathcal{R}_j$ definida en el producto cartesiano $\mathcal{D}_j \times \mathcal{D}'_j$.Sean $\mathcal{D} := \mathcal{D}_1 \times ... \times \mathcal{D}_p$ y $\mathcal{D}' := \mathcal{D}'_1 \times ... \times \mathcal{D}'_p$ los correspondientes productos cartesianos de los conjuntos de descripciones. La relación producto $\mathcal{R} = \mathcal{R}_1 \times ... \times \mathcal{R}_p$ definida en $\mathcal{D} \times \mathcal{D}'$ es la aplicación:*

$$\mathcal{R} : \mathcal{D} \times \mathcal{D}' \to \mathcal{L}$$
$$(d, d') \to \mathcal{R}(d, d') := [d\mathcal{R}d'] :=$$
$$g\left(\left\{\left[d_j \mathcal{R}_j d'_j\right], j = 1, ..., p\right\}\right) = \bigwedge_{j=1,...,p} \left[d_j \mathcal{R}_j d'_j\right]$$

que a cada par de descripciones $(d, d') \in \mathcal{D} \times \mathcal{D}'$, $d = (d_1, ..., d_p)$, $d' = \left(d'_1, ..., d'_p\right)$ le asocia un valor, denotado por $[d\mathcal{R}d']$. La aplicación g es una aplicación simétrica, que en general es el operador conjuntivo lógico estándar.

Definition 5. *Un objeto simbólico de tipo evento en E es una t-upla $(a, \mathcal{R}, d)$ donde:*

- *a es una función, denotada por* $a = [X\mathcal{R}d]$*, con* X *una variable simbólica con dominio* $\mathcal{Y}$ *definida por* $X : E \to \mathcal{D}$*,* $\mathcal{D}$ *un conjunto de descripciones de elementos de* E*, asociado al conjunto* $\mathcal{Y}$*. La función* $a : E \to \mathcal{L}$ *es tal que a cada elemento* $e \in E$ *le asocia el nivel de relación de su descripción en* $\mathcal{D}$ *(dada por* X*) con la descripción* d*.*
- $\mathcal{R}$ *es una relación de dominio definida en* $\mathcal{D} \times \{d\}$
- d *es una descripción de un conjunto de descripciones asociado al conjunto* $\mathcal{Y}$*.*

Definition 6. *Sea* $(a, \mathcal{R}, d)$ *con* $a = [X\mathcal{R}d]$*, un evento booleano definido en* E*. Se llama extensión del evento booleano a en* E *y se denota por* $Ext_E(a)$*, al subconjunto de elementos de* E *cuya descripción en* $\mathcal{D}$ *(dada por* X*) se relaciona con el evento a:*

$$Ext_E(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] = 1\}$$

Sea $(a, \mathcal{R}, d)$ *con* $a = [X\mathcal{R}d]$*, un evento definido en* E *y* $\delta \in [0, 1]$*. Se llama extensión de nivel* δ *del evento a en* E *y se denota por* $Ext_{E,\delta}(a)$*, al subconjunto de elementos de* E *cuya descripción en* $\mathcal{D}$ *(dada por* X*) tiene un nivel de relación con el evento a igual o superior a* δ *:*

$$Ext_{E,\delta}(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] \geq \delta\}$$

Existen objetos simbólicos tipo aserción, es decir referidos a varias variables. Se compone de varios eventos y está dotada de una función combinación de niveles de relación. Esta función combina los niveles de relación de cada uno de los eventos aplicados a un elemento del conjunto sobre el cual está definida la aserción.

3 Aplicación a la Encuesta Permanente de Hogares

Se aplican metodologías del Análisis Simbólico [6, 7, 8, 9],a datos de la Encuesta Permanente de Hogares (EPH), tercer trimestre de 2016, correspondiente al Gran San Juan. La EPH fue proporcionada por el Instituto Nacional de Estadísticas y Censos (INDEC). La encuesta consta de dos cuestionarios: uno, con datos de la vivienda y características del hogar y otro individual, con datos laborales, de ingresos, de educación y de migración de cada uno de los componentes del hogar.

Antes de comenzar el análisis se eliminaron las variables con información redundante, que tuvieran poca información, aquellas variables que dieran información muy detallada y se identificaron los registros que contenían información inconsistente a fin de eliminarlos.

El objetivo es analizar las ventajas del estudio de una gran base de datos mediante su transformación a una base de datos simbólicos.

Un primer análisis consiste en trabajar con la base de individuos y a partir de ella considerar al hogar como objeto simbólico a fin de describir las variables simbólicas de interés.

Table 1. Variables Simbólicas

Variable de Intervalo	Variable Modal
Edad	Relación de Parent.
Ingreso Total Fliar(ITF)	Estado
Ingreso Per Cápita Fliar (IPCF)	Estado Civil
Monto de Ingreso por Act.	Nivel de Educ.
Principal(MIAP)	Categoría de Act.
	Categoría de Inac.

La obtención de los objetos simbólicos (OS) se ha realizado mediante el software SODAS (Symbolic Official Data Analysis System). Este software provee muy buenas posibilidades de aplicación para la manipulación de bases de datos de estadísticas oficiales. A continuación se muestran e interpretan algunos resultados, salidas del software SODAS.

Variables simbólicas de Intervalo

Table 2. Distribución de las Variable Simbólica Edad

limits:0 - 94	class width: 9.4
class1	0.0456
class2	0.0970
class3	0.1359
class4	0.1755
class5	0.1369
class6	0.1076
class7	0.1100
class8	0.1237
class9	0.0566
class10	0.0112
Central tend.: 42.87	Disp.: 21.36

De acuerdo a la Tabla 2, se observa que la mayoría de los hogares tiene ocupantes con edades comprendidas entre 19 y 47 años. Es decir, se destacan hogares con ocupantes jóvenes. En menor proporción, hogares con adultos mayores entre 68 y 75 años. Hay pocos hogares con ancianos de más de 75 años. La edad promedio es de 42 años, con una dispersión de 21 años.

La distribución de ITF es muy asimétrica positiva, hay muchos hogares con ingresos por debajo de la media. La mayoría de los hogares tienen ingresos totales inferiores a $25665, siendo el intervalo modal el correspondiente a ingresos que varían de $9000 a $17300, aproximadamente. Los hogares con ingresos totales superiores a $33970 no llegan al 9%. El ingreso medio, de aproximadamente $16746 , no es representativo (ver Tabla 3).

Table 3. Distribución de las Variable Simbólica ITF

limits: 750 - 83803	class width: 8305.3
class1	0.2807
class2	0.3779
class3	0.1575
class4	0.1007
class5	0.0350
class6	0.0197
class7	0.0197
class8	0.0000
class9	0.0044
class10	0.0044
Central tend.:16746.35	Disp.:12531.02

Table 4. Distribución de las Variable Simbólica IPCF

limits:187.5 - 28500	class width:2831.2
class1	0.3001
class2	0.2973
class3	0.1772
class4	0.1094
class5	0.0503
class6	0.0284
class7	0.0241
class8	0.0044
class9	0.0044
class10	0.0044
Central tend.6056.70	Disp.:4693.85

Table 5. Distribución de las Variable Simbólica MIAP

limits:0 - 60000	class width:6000
class1	0.3854
class2	0.3843
class3	0.1491
class4	0.0726
class5	0.0057
class6	0.0008
class7	0.0008
class8	0.0006
class9	0.0003
class10	0.0003
Central tend.8651.40	Disp.: 5895.57

La distribución de IPCF muestra un comportamiento similar a ITF, hay muchos hogares con ingresos por debajo de la media de $6056. El 88,4% de los hogares tienen ingresos per cápita inferiores a $11500, aproximadamente, siendo los intervalos modales los correspondiente a ingresos per cápita menores a $5850, aproximadamente. Hay alrededor de un 10% de hogares con IPCF superiores a $13100 (Tabla 4).

En cuanto a la variable Monto de Ingreso de la Actividad Principal (sólo para hogares con integrantes ocupados) presenta mucha dispersión debido a valores de ingresos altos muy atípicos. Aproximadamente el 92 % de los hogares tiene ingreso de la actividad principal de sus integrantes inferiores a $ 18000 (Tabla 5).

Variables Modales

Las variables analizadas son:

- Estado Civil (CH07) 1 = unido 2 = casado 3 = separado/a ó divorciado/a 4 = viudo/a 5 = soltero/a
- Estado (Social) 1=Ocupado 2=Desocupado 3=Inactivo 4= Menor de 10 años
- Categoría Ocupacional (CAT-OCUP)(Para ocupados y desocupados con ocupación anterior) 1 = Patrón 2 = Cuenta propia 3 = Obrero o empleado, 4 = Trabajador familiar sin remuneración 9 = Ns./Nr. NA=No aplicable
- Categoría de Inactividad (CAT-INAC) 1 = Jubilado/ Pensionado 2 = Rentista 3 = Estudiante 4 = Ama de casa 5 = Menor de 6 años. 6 = Discapacitado 7 = Otros NA=No aplicable

Del análisis de la salida de SODAS, para el caso de variables modales, se observó que, en cuanto al estado civil, el 43% de hogares tienen integrantes solteros, mientras que un 27% aproximadamente tiene integrantes casados, el resto pertenecen a otra categoría de Estado Civil (Tabla 6).

Aproximadamente el 36% de hogares tiene integrantes ocupados y el 52% de hogares, integrantes inactivos. Hay un 28% de hogares donde la ocupación de sus integrantes es obrero o empleado, un 17% de hogares con integrantes inactivos estudiantes y un 29% de hogares con integrantes jubilados (Tabla 7).

Table 6. Capacidades de las Variables Modales CH07 y Estado

Variable CH07		Variable Estado	
Modalidad	Media	Modalidad	Media
1	0.1121	1	0.3606
2	0.2721	2	0.0162
3	0.0702	3	0.5235
4	0.1171	4	0.0997
5	0.4285		

Table 7. Capacidades de las Variables Modales Cat-Ocup y Cat-Inac

Variable Cat-Ocup		Variable Cat-Inac	
Modalidad	Media	Modalidad	Media
1	0.0150	1	0.2866
2	0.0714	2	0
3	0.2834	3	0.1670
4	0.0022	4	0.0788
9	0.0047	5	0.0613
NA	0.6232	6	0.0050
		7	0.0245
		NA	0.3768

Gráficos Comparativos

Con el fin de comparar el comportamiento de algunas variables, en relación con la situación laboral del encuestado, se las grafica para los grupos de Ocupados, Desocupados e Inactivos. Esta representación se efectúa considerando como objetos simbólicos las categorías de la variable Estado de Ocupación. Es decir, se trabaja sólo con cuatro objetos: Ocupados, Desocupados, Inactivos y No Aplicables (Niños menores de 10 aõs). La semántica utilizada es la de probabilidades basadas en la frecuencia.

Las variables analizadas, además de CH07, Estado, Cat-Ocup. y Cat-Inac son:

- Relación de Parentesco (CH03) 1 = Jefe/a 2 = Cónyuge/Pareja 3 = Hijo/Hijastro/a 4 = Yerno/Nuera 5 = Nieto/a 6 = Madre/Padre 7 = Suegro/a 8 = Hermano/a 9 = Otros Familiares 10 = No Familiares
- Sexo (CH04) 1 = varón 2 = mujer
- ¿Sabe leer y escribir? (CH09) 1 = Si 2 = No 3 = Menor de 2 años
- ¿Asiste o asistió a algún establecimiento educativo?(colegio, escuela, universidad)(CH10) 1 = Si, asiste 2 = No asiste, pero asistió 3 = Nunca asistió
- ¿Finalizó ese nivel?(CH13) 1 = Si 2 = No 9 = Ns./Nr.
- Nivel Educativo(NIVEL-ED) 1 = Primaria Incompleta(incluye educación especial) 2 = Primaria Completa 3 = Secundaria Incompleta 4 = Secundaria Completa 5 = Superior Universitaria Incompleta 6 = Superior Universitaria Completa 7 = Sin instrucción 9 = Ns./ Nr.
- ¿Cuánto hace que está buscando trabajo?(PP10A)(Sólo Desocupados) 1 = ...menos de 1 mes? 2 = ...de 1 a 3 meses? 3 = ...más de 3 a 6 meses? 4 = ...más de 6 a 12 meses? 5 = ...más de 1 año
- Ha trabajado alguna vez?(PP10D)(Sólo Desocupados) 1= Si 2= No
- Edad (CH06CAT): la variable Edad se categoriza en clases de amplitud 10 años.
- Ingreso Total Familiar Categorizado (ITF-CAT): la variable ITF se categoriza en clases de amplitud $10000.

La visualización de un objeto simbólico se hace mediante un gráfico llamado Zoom Star. Esta representación se basa en los diagramas de Kiviat donde cada eje representa una variable. En el mismo gráfico pueden representarse variables categóricas, de intervalo, con pesos, taxonomías, etc, sin sobrecargar el gráfico. SODAS permite dos tipos de representación, en $2D$ y $3D$, que muestran diferentes niveles de detalle. La representación en $2D$ permite una impresión global del objeto, mientras que la representación en $3D$ nos da información más detallada. En $2D$ los ejes están unidos por una línea que conecta los valores más frecuentes de cada variable. Si hubiera un empate del valor más frecuente en varias modalidades, la línea uniría las dos. Cuando existe una variable intervalo la línea se une a los límites mínimo y máximo y el área entera se rellena [10]. Mostramos la visualización $2D$ de los OS de interés:

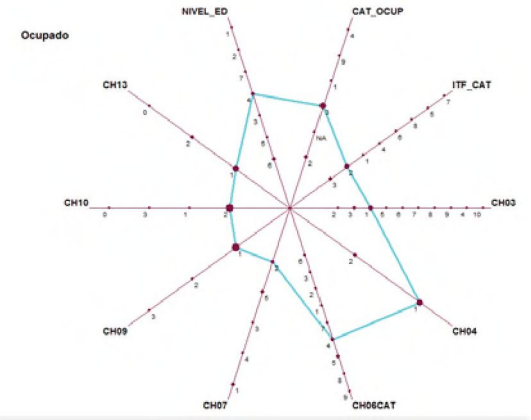

Fig. 1. Ocupados

El grupo de encuestados ocupados, Fig.1, se caracteriza (modos de cada variable) por haber asistido a un establecimiento educativo y finalizado ese nivel, teniendo un nivel de estudios secundario completo y, en menor proporción, universitario completo; su categoría ocupacional es obrero o empleado teniendo ingreso total familiar entre $10000 y $20000 y en menor proporción entre $20000 y $30000. Son jefes de hogar, varones, con edades entre 29 y 49 años, casados o solteros, casi en la misma proporción y todos saben leer y escribir.

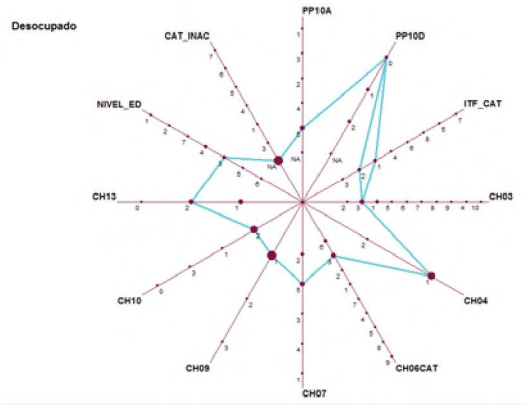

Fig. 2. Desocupados

El grupo de encuestados desocupados, Fig.2, se caracteriza por haber asistido a un establecimiento educativo y no haber finalizado ese nivel. Tienen un nivel de estudio secundario incompleto, han trabajado alguna vez y buscan trabajo hace más de un año, el ingreso total familiar de su hogar es menor a $20000. Son hijos, varones, con edades entre 29 y 39 años, solteros y todos saben leer y escribir.

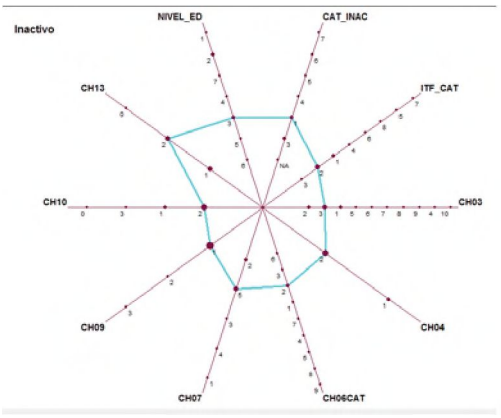

Fig. 3. Inactivos

El grupo de encuestados inactivos, Fig.3, se caracteriza por haber asistido a un establecimiento educativo y no haber finalizado ese nivel, teniendo un nivel de estudios secundario incompleto (el 35% son estudiantes y el 38% jubilados o pensionados). Tienen un ingreso total familiar inferior a $20000. Son hijos y en menor proporción jefes de hogar, varones, con edades entre 10 y 19 años, solteros, saben leer y escribir.

4 Conclusiones

Este trabajo muestra que los Datos Simbólicos son una herramienta de gran utilidad para el manejo y análisis de grandes volúmenes de datos.
Se destacan la posibilidad de trabajar con bases individuales relacionadas a fin de responder a diferentes propósitos de estudio, como también la generación de objetos simbólicos que son descriptos por variables simples, multivaluadas, probabilísticas y del tipo intervalo. Además, la teoría de objetos simbólicos posibilita trabajar con una población de interés específico. Por ejemplo, los integrantes del hogar de acuerdo a su Estado Social, permitiendo la comparación del comportamiento de diferentes variables en distintos grupos, donde cada grupo es un objeto simbólico de la tabla de datos simbólicos. También se destaca la posibilidad de la visualización exploratoria de los datos a través de gráficos $2D$ (o $3D$).

Existen numerosos análisis que pueden llevarse a cabo. En particular para la base analizada se pueden

construir objetos simbólicos tipo aserción obtenidos a partir de variables categóricas, por ejemplo combinando las modalidades de Nivel de Educación, Edad y Sexo. Para éstos objetos simbólicos se pueden aplicar métodos de clasificación no supervisada , a fin de detectar grupos homogéneos de integrantes de los hogares encuestados; respecto a variables de interés.

References

[1] Diday, E. (1987):Introduction a lapproche symbolique en analyse des données. Premiéres Journées Symbolique - Numérique. CEREMADE, Université Paris Dauphine, 21-56.

[2] Diday, E. (1988): The symbolic approach in clustering and related methods of data analysis : the basic choices. In: H.H. Bock (Ed.)

[3] Diday E. (1991): Des objets de lAnalyse des Données á ceux de lAnalyse des Connaissances in Induction symbolique et numérique. Y. Kodratoff and E. Diday edit. CEPADUESEDITIONS, Toulouse, France

[4] Diday E: (1995):Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations Researche 55, pp. 227-276.

[5] Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. Journal of the American Statistical Association, 98 (462), pp. 470487.

[6] Billard, L., Diday, E. (2007): Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.

[7] Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. WIREs Data Mining and Knowledge Discovery, Volume 4, Issue 4, July/August 2014, 281-295.

[8] Diday, E. An Introduction to Symbolic Data Analysis and the Sodas Software. University Paris,Dauphine.(2000)

[9] Diday, E. The state of the art in symbolic data analysis: overview and future. (2008)

[10] Bock, H.-H.; Diday, E. (2000): Analysis of Symbolic Data: Exploratory methods for extracting statistical information from complex data. Berlin-Heidelberg: Springer-Verlag.

[11] Bravo Llatas, M. Análisis de Segmentación en el análisis de datos Simbólicos.Tesis de la Universidad Complutense de Madrid, Facultad de Ciencias Matemáticas, Departamento de Estadística e Investigación Operativa. (2001)