Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

Extracting Information from Multivalued Surveys or from Very Extensive Data Sets by Symbolic Data Analysis

Edwin Diday*

Abstract

In many domains the units may be more complex than the standard ones due to the fact that they have internal variation and may be structured. Their description needs more complex data tables called "symbolic data tables" which are described in this paper. The need to extend standard data analysis methods (e.g., exploratory, clustering, factor analysis, discriminant analysis) to the symbolic data table is growing due to the need of getting more accurate information and to summarize extensive data sets. We call "Symbolic Data Analysis" (SDA) the extension of standard Data Analysis to such tables. "Symbolic objects" describe in an explanatory way classes of units described by symbolic data. They constitute one of the main output of a SDA. A symbolic object is "complete" if its "extent" covers exactly the class that describes it. An illustrative example concerning the pollution of several locations at different time points is given. This example yields to stochastic concept lattices which are finally introduced.

1 Introduction

In order to obtain more accurate surveys, individuals may be allowed to give multivalued answers. For instance, the answer to a query may be a set of categorical values which expresses the doubt of an individual answer.

All around the world, huge data sets are recorded in Official Statistics as well as in Companies. Summarizing such information in smaller sets of new « statistical units » is a question of increasing importance. In reducing the data set and losing the least information possible, these statistical units yield more complex data tables called « symbolic data tables », because the cells of such data tables may contain not only

* University Paris IX Dauphine, Place de Marechal de Lottre de Tassigny, 75775 Paris CEDEX 16, France

single numerical or categorical values, but much more complex information, such as: subsets of categorical variable values, intervals of ordinal variable values, histograms, probability distributions and dependencies, and need rules to be specified. These new statistical units are called « symbolic objects » and there is a need for an extension of standard Data Analysis to such objects, called « Symbolic Data Analysis » .

2 Main input of Symbolic Data Analysis algorithms: « symbolic data tables »

Columns of the initial data table are « variables » and rows are « symbolic descriptions ». Each cell of this « symbolic data table » may contain data of different types:

a. Single quantitative value : for instance, if « weight » is a variable and w is a unit : weight(w)=3.5.
b. Single categorical value: for instance, Town(w)= London.
c. Multivalued: for instance, in the quantitative case: weight(w)={3.5, 2.1, 5}which means that the weight of w may be 3.5 or 2.1 or 5. In the categorical case, color(w) = {blue, red, yellow} means that the color of w may be blue or red or yellow. Notice that (a) and (b) are special cases of (c).
d. Interval: for instance weight(w)=[3, 5], which means that the weight of w varies in the interval [3, 5].
e. Multivalued with weights: for instance a histogram or a membership function (notice that (a) and (b) and (c) are special cases of (e) when the weights are equal to 1).

Variables may be:

f. Taxonomic: for instance, « the color is considered to be light if it is yellow, white or pink ».
g. Hierarchically dependent: for instance, we may describe the kind of computer of a company only if it has a computer, hence the variable "does the company have computers? " and the variable " kind of computer" are hierarchically linked.
h. With logical dependencies: for instance, « if age(w) is less than 2 months then weight(w) is less than 10 ».

3 Main output of Symbolic Data Analysis algorithms: complete symbolic objects

Let Ω be a set of units called « individuals » , D a set of descriptions, « y » a mapping (called attribute or variable) defined from Ω into D, which associates with each $w \in \Omega$ a description $d \in D$ from a given symbolic data table. We denote by R, a « comparison » operator between two descriptions such that [d' R d] $\in$ L where L = {true, false} or L= [0,1]. For instance $R \in \{=, \neq, \equiv, \leq, \geq, \subseteq,, \supseteq, \notin, \in$, an implication, a kind of matching, ...). A set of coherent descriptions (for instance, the description: sex = male and number of deliveries = 1 is not coherent), constitutes the set of objects to which any symbolic data analysis algorithm applies. That is why, we use « object » to denote any coherent description . If it is the description of an individual, we call it an « individual object ». It may be also the description of a class of individuals, of a scenario, of a strategy, etc. In the case of a survey, an individual object is a coherent answer of an individual to the set of queries in the survey.

A « symbolic object » is defined by an object, and a way of comparing it to individual objects is defined by a recognition mapping. The advantages of «symbolic objects» are of at least two kinds. First, they give a summary of the initial symbolic data table in an explanatory way, (i.e. close to the initial language of the user) by expressing descriptions based on the marginal distributions of the initial variables. Secondly, by being independent of the initial symbolic data table, they are able to identify any matching individual described in any data table. More formally, their definition is:

Definition of a symbolic object
A symbolic object is a triple s = (a, R, d) where R is a comparison operator, d is a description and « a » is defined from Ω in L such that a(w) = [y(w) R d].

There are two kinds of symbolic objects:

- « boolean symbolic objects » if [y(w) R d] $\in$ L = {true, false}. In this case, the y(w) are of type (a) to (d), defined in section 1.
  Example: d = {red, blue, yellow}, color(w) = {red, yellow}, R= « $\subseteq$ »,
  a(w) = [color(w) $\subseteq$ d} = true.
- « modal symbolic objects » if [y(w) R d] $\in$ L = [0,1]. In this case, the y(w) are of type (e). An example of choice for R is given hereunder.

Extent of a symbolic object s: in the boolean case, it is defined by Ext(s) = {w $\in \Omega$ / a(w) = true}. In the modal case, given a threshold α, it is defined by Ext(s) = {w $\in$ Ω / a(w) $\geq \alpha$}.

Tools for symbolic objects: tools between symbolic objects (Diday, 1995) may be needed, such as similarities, matching, merging by generalisation where a « t-norm » or a « t-conorm » (Schweizer and Sklar, 1983) or « capacities » (Diday and Emilion, 1997) may be used, splitting by specialization (Ciampi et al., 1996) . Let T be a merging operator, which associates a description with a set of descriptions (for instance, the « sup » or the « inf » of such a set). If the choice of R and T is coherent, it may be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Polaillon and Diday, 1997), where the vertices are closed sets defined by « complete symbolic objects ». More precisely, the associated Galois correspondence (see e.g. Wille, 1983) is defined by two mappings:

- F: from $P(\Omega)$ (the power set of Ω) into S (the set of symbolic objects) such that $F(C)= s$ where $s = (a, R,d)$ is defined by $d = T_{c\in C}\, y(c)$ and so $a(w) = [y(w)\ R\ T_{c\in C}\, y(c)]$, for a given R.
  For example, if $y(u) = \{pink, blue\}$, $C = \{c, c'\}$, $y(c)= \{pink, red\}$, $y(c')= \{blue,red\}$, and if $T(\{pink, red\},\{blue,red\}\})=\{pink, red, blue\}$)and $R \equiv « \subseteq »$, then $a(u)=$ true and $u\in$ Ext (s).
- G: from S into $P(\Omega)$ such that: $G(s) =$ Ext (s).

Complete symbolic object: A symbolic object s is a « complete symbolic object » iff $F(G(s)) = s$.
These objects may be selected from the Galois lattice but also if the lattice is too large, from a partitioning, a hierarchical or a pyramidal clustering, from the most contributive individuals to a factorial axis, from a decision tree, etc. extended to have symbolic objects as input and giving complete symbolic objects as output.

Syntax of symbolic objects: if the initial data table contains p variables we denote $y(w) = (y_1(w),..., y_p(w))$, $D = (D_1,...,D_p)$, $d\in D$: $d=(d_1,..., d_p)$ and $R = (R_1,...,R_p)$. Then an « assertion » is a special case of a symbolic object defined by $s = (a,R,d)$ and written as follows: $a(w) = \wedge_{i=1,p}\, [\, y_i(w)\ R_i\ d_i]$.

Individual symbolic objects, first and second order symbolic objects: any row describing an individual « u » of a symbolic data table induces an assertion called an « individual symbolic object » by setting: $a(w) = \wedge_{i=1,p}\, [\, y_i(w)\ R_i\ y_i(u)\ ]$ where $R_i =$ « = » . Its extent is the set of individuals with the description defined in the symbolic data table. Hence, if all the individuals have different descriptions, there is a bijection between the set of individuals defined in a symbolic data table, and their associated individual objects or their associated individual symbolic objects. That is why sometimes a row of a symbolic data table is called a « symbolic object ». We distinguish « first order symbolic objects » (associated with individuals) from « second order symbolic objects » which are associated with the description (obtained by generalization) of a class of symbolic objects. For example, a second

order boolean assertion object where SPC means « socio-professional-category » is: $a(w) = [age(w) \subseteq \{12, 20, 28\}] \wedge [SPC(w) \subseteq \{employee, worker\}]$. If the individual u is described in the initial symbolic data table by $y_1(u)=\{12, 20\}$ and $y_2(u) = \{employee\}$ then its associated individual symbolic object is:

$u^s(w) = [age(w) = \{12, 20\}] \wedge [SPC(w) = \{employee\}]$. Now,

$a(u) = [\{12, 20\} \subseteq \{12, 20, 28\}] \wedge [\{employee\} \subseteq \{employee, worker\}]$ and therefore a(u)=true.

If the variables are multivalued and weighted, an example of an assertion is:

$a(w) = [age(w)\ R_1\ \{(0.2)12, (0.8)[20, 28]\}] \wedge [SPC(w)\ R_2\ \{(0.5)employee, (0.5)worker\}]$ where for instance the « matching of two probability distributions » is defined for two discrete distributions r and q of k values by:

$r\ R_i\ q = \sum_{j=1,k} r_j\ q_j\ e^{(r_j - \min(r_j, q_j))}$.

4 Surce of Symbolic Data

Symbolic data arise from many sources where there is a need to summarize huge sets of data tables. They result from the probability distribution, the percentile or the range of any random variable associated with each cell of such data tables. They result also, from Relational Data Bases (in merging several relations or summarizing the answer to a query), from surveys giving the possibility of multivalued answers, from Data Analysis (factorial analysis, clustering, neural networks,...) of standard data table, from expert knowledge (scenario of traffic accidents, type of unemployed,...), from time series (in describing intervals of time), from confidential data (in order to hide the initial data by less accuracy), etc.

Table 1: The initial data table

	Y_1: POLLUTION								Y_2: ENVIRONMENT							
	y_1				y_p				z_1			z_p				
EIFFEL TOWER (T_1)																
.....																
EIFFEL TOWER (T_n)																
.....																
DAUPHINE (T_1)																
.....																
DAUPHINE (T_k)																
.....																

5 An illustrative example

The set of individuals Ω in this example, is a set of locations in Paris (Eiffel Tower, Dauphine University, Place de l'Etoile,...) considered at different times ($T_1, T_2, ...$). We have two sets of variables: the first, $\mathbf{Y_1}$, concerns pollution (rate of lead, of CO_2,....). The second, $\mathbf{Y_2}$, is a set of environment variables (density of cars, direction of the wind, ...). Hence,we have a huge initial data table which is described in Table 1.

The question is expressed in the following way: find classes of locations which have the same environment conditions, describe them in a humanly comprehensible language, organize them in terms of this language, be able to allocate a new location to these classes, and study the evolution of these classes when the number of observations increases.

Table 2: Using a clustering algorithm, each location x time is assigned to a class

	POLLUTION	ENVIRONMENT
EIFFEL(T_1)	POL 9	ENV 4
EIFFEL(T_2)	POL 5	ENV 7
.....		
EIFFEL(T_n)	POL 11	ENV 9
.....		
DAUPHINE(T_1)	POL 5	ENV 9
.....		
DAUPHINE(T_k)	POL 6	ENV 7

We propose the following method:

1. Partitioning (using a good clustering algorithm) the locations x times described by the pollution variables. The K classes of this partition are denoted: POL 1, POL 2, ... , POL K.
2. Partitioning (with the same algorithm) the locations x times described by the environmental conditions. We obtain L classes of locations x times denoted: ENV 1, ENV 2, ... , ENV L.
3. We build data table 2, which associates with each location at each time a class characterized by the pollution variables and a class characterized by the environment variables.

4. We build the histogram of pollution behaviour for each location, for each class of environmental conditions. For instance, in the following data table 3 , we have K=4 , L = 3 . The histogram contained in the first cell defines for each class of pollution (POL 1, POL 2, POL 3, POL 4), the frequency of the Eiffel Tower x time cases, in the first environmental class: ENV 1. In the graphical representation of each histogram, each class POL i is associated with an interval (see Figure 1).

Table 3: Location behavior in three classes of environmental conditions.

	ENV 1	ENV 2	ENV 3
EIFFEL TOWER			
DAUPHINE UNIVERSITY			
PLACE DE LA CONCORDE			

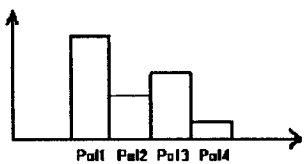

Figure 1: The level associated with the class Pol i of the histogram associated with a location and an environment class Env j in Table 3, is the number of times (obtained from Tables 1 and 2) when this location is simultaneously in the class Pol i and in the class Env j.

Having such a « context » defined by the triple: locations (the objects), environment classes (the variables), histograms (the variable values), we are in the « symbolic data analysis framework » where Ω is a set of locations, $y(w) = (h_1(w), h_2(w), h_3(w))$ is an individual object which describes the individual w, by three histograms, where h_i corresponds to the variable y_i. In other words, each individual object is a row of the preceding data table of histograms. In order to describe a class C of locations we can use for instance a t-norm denoted T such that:

$$T_{c \in C}\, y(c) = (\mathrm{Max}_{c \in C}\, h_1(c), \mathrm{Max}_{c \in C}\, h_2(c), \mathrm{Max}_{c \in C}\, h_3(c)).$$

The comparison operator is then $R_i = « \leq »$, in order to be coherent with T. Hence, by using the notations given in 2, we have $F(C) = s$ where $s = (a, R, d)$ is defined by

$$d = T_{c \in C}\, y(c) \text{ and } a(w) = \wedge_{i=1,p} [\, y_i(w)\ R_i\ T_{c \in C}\, y_i(c)].$$

In other words, $a(w) = \wedge_{i=1,p} [\, h_i(w) \leq \mathrm{Max}_{c \in C}\, h_i(c)]$. It follows that the extension of s is

$$G(s) = \{w / a(w) = \text{true}\} = \{ w / \ \forall i, h_i(w) \leq \mathrm{Max}_{c \in C}\, h_i(c) \}.$$

As in this case R and T are coherent, F and G constitute a Galois correspondence. Each node of the associated concept lattice corresponds to a complete symbolic object s: $F(G(s)) = s$.

This lattice defines all the complete symbolic objects associated with F and G.

6 Stochastic concept lattices

6.1 Galois lattice of probability distributions

The basic framework of stochastic concept lattices is the following:

Ω is a finite set of individuals (for example, the set of « location x time » described by a pollution class for each environment class). P is a probability measure defined on $P(\Omega)$ the power set of Ω. J is a finite set of indices (the variables: the environment classes). O_j is a finite set (the pollution classes). $\{(X_{i,j})\ j \in J\}$ is a family of random variables from Ω to O_j. For example, $X_{i,j}(w)$ is the « class pollution » of the i-th location at time w in the class environment j.

Let $L_{i,j}$ be the probability distribution of $X_{i,j}$. We define the following symbolic object :

$F(C) = (a, R, d)$ where $R_i = « \in »$, $d_C = [(\text{Min}_{c \in C} L_{c,j})_j, (\text{Max}_{i \in C} L_{c,j})_j]$, and $a(w) = \wedge_{i=1,p} [L_{w,j} \in [(\text{Min}_{c \in C} L_{c,j})_j, (\text{Max}_{c \in C} L_{c,j})_j]$.

In other words:
$a(w) = \wedge_{i=1,p} [[(\text{Min}_{c \in C} L_{c,j})_j \leq L_{w,j} \leq (\text{Max}_{c \in C} L_{c,j})_j]$.

The following results may be shown (Diday and Emilion, 1997):

Theorem 1:
F and G are decreasing, h = GoF and k=FoG are increasing extensive and idempotent, so that we get a concept lattice. In the case of a binary random variable we get the usual binary concepts lattice.

6.2 Galois lattice of histograms

Suppose that the $L_{i,j}$ are unknown but that we have at our disposal frequency histograms $H_{i,j}$. We can define the same kind of symbolic objects by using $H_{i,j}$ instead of $L_{i,j}$. Then we have:

Theorem 2:
F and G yield a concept lattice and if morover the frequency histograms converge, then the lattices defined by these histograms converge to that of Theorem 1.

6.3 Galois lattice of support measures

The closed support of the measure L_{ij} is denoted by S_{ij} .We suppose that we have got some observations of the random variables $X_{i,j}$ such that the sequence of vectors $\{X_{i,j}{}^{(n)} \; i \in I\}_n$ is independent and distributed like the vector $\{X_{i,j} \; i \in I\}$.

The mappings F and G are defined as follows:

$F(C) = s$ such that $s = (a, R, d)$ with
$R = « \subseteq »$, $d = (\cap_{i \in C} S_{ij})_j$ and $a(w) = \wedge_{i=1,p} [S_{wj} \subseteq \cap_{i \in C} S_{ij}]$.
$G(s) = \{w / S_{wj} \subseteq V_j \text{ for all } j\}$ if $d = (V_j)_j$.

We have then the following result:

Theorem 3:
As n goes to infinity, the step n - lattices converge to the Galois lattice induced by F and G.

As a consequence, Theorems 2 and 3 show that as the knowledge of the individuals increases the concepts focus and converge.

7 An example of a lattice

The initial symbolic data table is given in Table 4. It may be obtained in a stochastic lattice context, from the support of each probabilty distribution L_{ij} or from a histogram H_{ij} or from a given percentile.

Table 4: The initial symbolic data table

	y1	*y2*	*y3*
1	a,b	∅	g
2	∅	∅	g,h
3	c	e,f	g,h,i
4	a,b,c	e	h

With $T \equiv \cup$ and $R \equiv \subseteq$, the complete symbolic objects denoted $s_i = (a_i, \subseteq, d_i)$ and their extension obtained by using an extension of the Chein algorithm (Diday, 1996), Pollaillon and Diday, 1997) are the following:

$a_1(w) = [y_1(w) \subseteq O_1] \wedge [y_2(w) \subseteq O_2] \wedge [y_3(w) \subseteq O_3]$, $Ext(s_1) = \{1,2,3,4\}$
$a_2(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_2) = \{1, 2, 4\}$
$a_3(w) = [y_1(w) \subseteq \{c\}]$, $Ext(s_3) = \{2, 3\}$
$a_4(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \emptyset] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_4) = \{1,2\}$
$a_5(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{h\}]$, $Ext(s_5) = \{4\}$
$a_6(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \emptyset] \wedge [y_3(w) \subseteq \{g\}]$, $Ext(s_6) = \{1\}$
$a_7(w) = [y_1(w) = \{\emptyset\}] \wedge [y_2(w) = \emptyset] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_7) = \{2\}$
$a_8(w) = [y_1(w) = \emptyset] \wedge [y_2(w) = \emptyset] \wedge [y_3(w) = \emptyset]$, $Ext(s_8) = \{\emptyset\}$

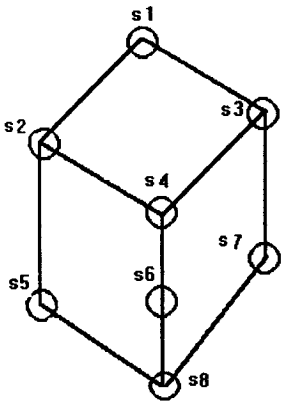

Figure2: Galois Lattice of the symbolic objects defined in Table 1.

8 Conclusion

A general aim of a symbolic data analysis may be stated in the following way: having as input: (Ω, D, y, T, R), find complete symbolic objects. But, as in statistics, the underlying lattice often becomes too large, and other methods which also provide symbolic objects have to be used. The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data tables is growing due to the expansion of information technology. This need has led to a European Community project called SODAS (Hebrail, 1996) for a « Symbolic Official Data Analysis System» in which 15 institutions in 9 European countries are involved.

References

[1] Ciampi A., Diday E., Lebbe J., Périnel E., and Vigne R. (1996): Recursive partition with probabilistically imprecise data. In: E. Diday, Y. Lechevallier, and O. Opitz (Eds): *Proceed. of OSDA '95* (Ordinal and Symbolic Data Analysis). Springer Verlag, Heidelberg, 201-214.

[2] Diday E. and Emilion R. (1996): Lattices and capacities in analysis of probabilist objects. In: E. Diday, Y. Lechevallier, and O. Opitz (Eds): *Proceed. of OSDA '95* (Ordinal and Symbolic Data Analysis). Springer Verlag, Heidelberg, 13-30.

[3] Diday E. and Emilion R. (1997): Treillis de Galois maximaux et capacités de Choquet. *Compte Rendu à l'Académie des Sciences.* Paris, T. 324, Sériel. Elsevier Academic.

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

[5] Diday E. (1996): Une introduction à l'analyse des donnees symboliques. *Proc. SFC*, Vannes, France.

[6] Hebrail G. (1996): SODAS (Symbolic Official Data Analysis System). *Proceedings of IFCS'96*, Kobe, Japan. Springer Verlag.

[7] Pollaillon G. and Diday, E. (1997): *Galois Lattices of Symbolic Objects.* Rapport du Ceremade University Paris9- Dauphine (February).

[8] Schweizer B. and Sklar A. (1983): *Probabilist Metric Spaces.* Elsever North-Holland, New-York.

[9] Wille R. (1983): Subdirect decomposition of concept lattices. *Algebra Universalis.*