FROM DATA MINING TO KNOWLEDGE MINING:
SYMBOLIC DATA ANALYSIS
AND THE SODAS SOFTWARE

E. Diday
University of Paris IX Dauphine
and INRIA

AIM

FROM HUDGE DATA IN AN ECONOMIC WAY

- Extract new knowledge
- Summarize
- Concatenate
- Solve confidentiality
- Explain correlation

HOW? By working on HIGHER LEVEL UNITS called CONCEPTS necessary described by more complex data extending Data Mining to Knowledge Mining.
OUTLINE

1) THE MAIN IDEA: FIRST AND SECOND ORDER OBJECTS.
2) THE INPUT OF A SYMBOLIC DATA ANALYSIS: SYMBOLIC DATA TABLE.
3) SOURCE OF SYMBOLIC DATA: FROM DATA BASES, FROM CATEGORICAL VARIABLES.
4) MAIN OUTPUT OF SYMBOLIC DATA ANALYSIS ALGORITHMS: SYMBOLIC DESCRIPTIONS AND SYMBOLIC OBJECTS.
5) THE MAIN STEPS OF A SDA.
6) TOOLS OF SYMBOLIC DATA ANALYSIS
7) SYNTHETICAL VIEW OF THE SODAS PROJECT

THE MAIN IDEA:
FIRST AND SECOND ORDER OBJECTS

THE ARISTOTLE ORGANON (IV B.C.) CLEARLY DISTINGUISHES "FIRST ORDER OBJECTS" (AS THIS HORSE OR THIS MAN) CONSIDERED AS A UNIT DESCRIBING AN INDIVIDUAL OF THE WORLD , FROM "SECOND ORDER OBJECTS" (AS A HORSE OR A MAN) ALSO TAKEN AS A UNIT DESCRIBING A CLASS OF INDIVIDUALS.
FOUR IDEAS

1) ONLY TWO LEVELS OF UNITS:
   First level: Individuals
   Second level: concepts

2) THE CONCEPTS CAN BE CONSIDERED AS NEW INDIVIDUALS OF HIGHER LEVEL.

3) A CONCEPT IS DESCRIBED BY USING THE DESCRIPTION OF A CLASS OF INDIVIDUALS OF ITS EXTENT.

4) THE DESCRIPTION OF A CONCEPT MUST EXPRESS THE VARIATION OF THE INDIVIDUALS OF ITS EXTENT

FROM FIRST ORDER OBJECTS TO SECOND ORDER OBJECTS IN OFFICIAL STATISTICS

<table>
<thead>
<tr>
<th>Units</th>
<th>Classes</th>
<th>Descr. Var. of the Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case #</td>
<td>Region</td>
<td>Bedroom</td>
</tr>
<tr>
<td>11401</td>
<td>Northern Metropolitan</td>
<td>2</td>
</tr>
<tr>
<td>11402</td>
<td>Northern Metropolitan</td>
<td>2</td>
</tr>
<tr>
<td>11403</td>
<td>Northern Metropolitan</td>
<td>1</td>
</tr>
<tr>
<td>12315</td>
<td>East Anglia</td>
<td>1</td>
</tr>
<tr>
<td>12316</td>
<td>East Anglia</td>
<td>2</td>
</tr>
<tr>
<td>14524</td>
<td>Greater London</td>
<td>1</td>
</tr>
</tbody>
</table>
FROM FIRST ORDER OBJECTS TO SECOND ORDER OBJECTS
IN OFFICIAL STATISTICS

<table>
<thead>
<tr>
<th>Classes</th>
<th>Descriptive variable of the units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>Bedroom</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>2</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>2</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>1</td>
</tr>
<tr>
<td>East-anglia</td>
<td>1</td>
</tr>
<tr>
<td>East-anglia</td>
<td>2</td>
</tr>
<tr>
<td>East-anglia</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classes</th>
<th>Descriptive variables of the classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>Bedroom</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>(2/3) 2, (1/3) 1</td>
</tr>
<tr>
<td>East-anglia</td>
<td>(2/3) 1, (1/3) 2</td>
</tr>
</tbody>
</table>

Schweitzer (1984): "Distributions are the numbers of the future"

MORE GENERALLY, WHAT IS THE INPUT OF A SYMBOLIC DATA ANALYSIS?
SYMBOLIC DATA TABLE
+ BACKGROUND KNOWLEDGE

SYMBOLIC DATA TABLE:
THE CELLS CAN CONTAIN:
- SEVERAL QUALITATIVE OR QUANTITATIVE WEIGHTED VALUES
- INTERVALS
- HISTOGRAMS
EXAMPLE OF SYMBOLIC DATA TABLE

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>WEIGHT</th>
<th>TOWN</th>
<th>COLOUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRODUCT 1</td>
<td>3.5</td>
<td>Londres</td>
<td>(red, white, yellow)</td>
</tr>
<tr>
<td>PRODUCT 2</td>
<td>[ 3, 8 ]</td>
<td>Paris, Londres</td>
<td></td>
</tr>
<tr>
<td>PRODUCT 3</td>
<td>[3.1, 4.6, 7.2]</td>
<td></td>
<td>(0.3 red, 0.7 green)</td>
</tr>
<tr>
<td>PRODUCT 4</td>
<td>[(0.4) [2,3], (0.6) [3, 8]]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

THE CELLS CAN CONTAIN:

- SEVERAL QUALITATIVE OR QUANTITATIVE WEIGHTED VALUES
- INTERVALS
- HISTOGRAMS

SOME BACKGROUND KNOWLEDGE CAN BE GIVEN

VARIABLES CAN BE:

- TAXONOMIC: « A SOCIO-ECONOMIC GROUP IS CONSIDERED TO BE "SELF-EMPLOYED" IF IT IS "PROFESSIONAL SELF-EMPLOYED" OR "OWN ACCOUNT NON-PROFESSIONAL". »
- HIERARCHICALLY DEPENDENT:
- WITH LOGICAL DEPENDENCIES:
  « IF AGE(W) IS LESS THAN 2 MONTHS THEN HEIGHT(W) IS LESS THAN 10 ».
SOURCE OF SYMBOLIC DATA

FROM CATEGORICAL VARIABLES:
- GIVEN (AS « TYPE OF EMPLOYMENT »)
- OBTAINED BY CLUSTERING.

FROM DATA BASES: QUERY CREATING A NEW CATEGORICAL VARIABLE: cartesian prod

FROM EXPERT: NATIVE SYMBOLIC DATA:
Scenario of road accidents, species of insects

FROM CONFIDENTIAL DATA
IN ORDER TO HIDE THE INITIAL DATA BY LESS ACCURACY

FROM STOCHASTIC DATA TABLE:
THE PROBABILITY DISTRIBUTION, THE HISTOGRAM THE PERCENTILES OR THE RANGE OF ANY RANDOM VARIABLE ASSOCIATED TO EACH CELL OF A DATA TABLE

EXAMPLE

<table>
<thead>
<tr>
<th></th>
<th>Mathematics</th>
<th>Physics</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>XM</td>
<td>Xp</td>
<td>XI</td>
</tr>
<tr>
<td>Paul</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

XM is the random variable which associates to each exam of TOM his mark in mathematics.
From XM several kinds of symbolic objects can be defined by using in each cell:
- Probability distr.
- Histograms
- Inter-quartile intervals
FROM TIME SERIES

- IN DESCRIBING INTERVALS OF TIME:
  (the variation of the values each week)

- IN DESCRIBING A TIME SERIES BY THE HISTOGRAM OF ITS VALUES.

FROM RULES

Example: Districts description

<table>
<thead>
<tr>
<th>Districts</th>
<th>male-full-time-employee %</th>
<th>male-part-time-employee %</th>
<th>male-self-employed %</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>8%</td>
<td>5%</td>
<td>20%</td>
<td>Z1%</td>
</tr>
<tr>
<td>D2</td>
<td>12%</td>
<td>9%</td>
<td>15%</td>
<td>Z2%</td>
</tr>
<tr>
<td>Dn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example of Rule defining class

• **R1**: male-full-time-employee%(X, low) \(\land\) male-part-time-employee%(X, low) \(\land\) neighbor(X, Y) \(\land\) comm-activities(Y, high)

\[\rightarrow\] male-self-employed%(X, high)

• 70 districts X satisfy the rule: the low percentage of full-time and part-time male employees in district X adjacent neighbor of Y, with many commercial activities, implies a high percentage of self-employed males in X.

### Districts description of rules

<table>
<thead>
<tr>
<th>Descript of rules extent</th>
<th>male-full-time-employee%</th>
<th>male-part-time-employee%</th>
<th>male-self-employed%</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>[8%, 12%]</td>
<td>[5%, 9%]</td>
<td>[20%, 15%]</td>
<td>[Z1%, Z2]</td>
</tr>
<tr>
<td>R2</td>
<td>[2%, 6%]</td>
<td>[4%, 8%]</td>
<td>[18%, 14%]</td>
<td>…..</td>
</tr>
<tr>
<td>Rn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MAIN OUTPUT OF SYMBOLIC DATA ANALYSIS ALGORITHMS:
SYMBOLIC DESCRIPTIONS
SYMBOLIC OBJECTS.

SYMBOLIC DESCRIPTIONS

<table>
<thead>
<tr>
<th>Description</th>
<th>AGE</th>
<th>SPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>{12,20,28}</td>
<td>{employee, worker}</td>
</tr>
<tr>
<td>D2</td>
<td>[5, 33]</td>
<td>{teacher, countryman}</td>
</tr>
</tbody>
</table>

CONCEPTS ARE MODELED BY SYMBOLIC OBJECTS

WHATS A CONCEPT?
A CONCEPT IS DEFINED BY AN
* INTENT : ITS CHARACTERISTIC PROPERTIES
* EXTENT: THE SET OF INDIVIDUALS WHICH SATISFY THESE PROPERTIES

LIKE OUR MIND, SYMBOLIC OBJECTS MODEL CONCEPTS BY AN INTENT AND A WAY OF COMPUTING THE EXTENT
SYMBOLIC OBJECT

It’s an animal(w) = 0.99 yes

S = (a, R, d_C)  \iff  a(w) = \{y(w)Rd_C\}

TWO KINDS OF SYMBOLIC OBJECTS

BOOLEAN SYMBOLIC OBJECTS

S = (a, R, d_1)

d_1= \{12, 20, 28\} x \{employee, worker\}

R = (\subseteq, \subseteq),

a(w) = \{age(w) \subseteq \{12, 20, 28\}\} \land [SPC(w) \subseteq \{employee, worker\}]

a(w) \in \{TRUE, FALSE\}. 
S = (a, R, d):
\[ a(w) = \{\text{age}(w) R_1 \{(0.2)12, (0.8) [20,28]\}\} \land \\{\text{SPC}(w) R_2 \{(0.4)\text{employee}, (0.6)\text{worker}\}\} \]
\[ a(w) \in [0,1]. \]

First approach: simple or flexible matching
\[ R = (R_1, R_2) : r \leq r_i q = \sum_{j=1,k} r_j q_j e^{(r_j - \min(r_j, q_j))}. \]

Second approach:
Probabilistic: if dependencies, copulas, derivation of the joint distribution, transforming the joint density in [0,1].

THE MEMBERSHIP FUNCTION « a » MODAL CASE

EXTENT OF A SYMBOLIC OBJECT S:

BOOLEAN CASE:
\[ \text{EXT}(s) = \{W \in \Omega / a(W) = \text{TRUE}\}. \]

MODAL CASE
\[ \text{EXT}_\alpha (S) = \text{EXTENT}_\alpha (a) = \{W \in \Omega / a(W) \geq \alpha\}. \]
AN OVERVIEW ON THE SODAS SOFTWARE
THE MAIN STEPS FOR A SYMBOLIC DATA ANALYSIS IN SODAS

. PUT THE DATA IN A RELATIONAL DATA BASE (Oracle, Access, ...)

. DEFINE A CONTEXT BY GIVING
  * The Units (Individuals, Households, ...)
  * The Classes (Regions, Socio-economics groups, ...)
  * The descriptive variables of the units

. BUILD A SYMBOLIC DATA TABLE WHERE
  * The units are the preceding classes
  * The descriptions of each class is obtained by a generalization process applied to its members

APPLY
SYMBOLIC DATA ANALYSIS TOOLS

- Correlation, Mean, Mean Square
- Histogram of a symbolic variable
- Dissimilarities between symbolic descriptions
- Clustering of symbolic descriptions
- Principal component Analysis
- Decision Tree
- Graphical visualisation of Symbolic Objects
SODAS Software

Menu

Methods

Sds file

Report

Graphs

Concatenation of summarized data from several populations

Join two or more sets of SO based on different underlying populations
DE CARVALHO’S DISSIMILARITY MEASURES

A straightforward extension of similarity measures for classical data matrices with nominal variables.

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Disagreement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>$\alpha = \mu(A_j \cap B_j)$</td>
<td>$\beta = \mu(A_j \cap c(B_j))$</td>
</tr>
<tr>
<td>Disagreement</td>
<td>$\chi = \mu(c(A_j) \cap B_j)$</td>
<td>$\delta = \mu(c(A_j) \cap c(B_j))$</td>
</tr>
<tr>
<td>Total</td>
<td>$\mu(B_j)$</td>
<td>$\mu(c(B_j))$</td>
</tr>
</tbody>
</table>

where $\mu(V_j)$ is either the cardinality of the set $V_j$ (if $Y_j$ is a nominal variable) or the length of the interval $V_j$ (if $Y_j$ is a continuous variable).

Five different similarity measures $s_i$, $i = 1, ..., 5$, are defined:

<table>
<thead>
<tr>
<th>$s_i$</th>
<th>Comparison Function</th>
<th>Range</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$\alpha/(\alpha + \beta + \chi)$</td>
<td>$[0,1]$</td>
<td>metric</td>
</tr>
<tr>
<td>$s_2$</td>
<td>$2\alpha/(2\alpha + \beta + \chi)$</td>
<td>$[0,1]$</td>
<td>semi metric</td>
</tr>
<tr>
<td>$s_3$</td>
<td>$\alpha/(\alpha + 2\beta + 2\chi)$</td>
<td>$[0,1]$</td>
<td>metric</td>
</tr>
<tr>
<td>$s_4$</td>
<td>$\chi/[(\alpha + \beta) + \chi/(\alpha + \chi)]$</td>
<td>$[0,1]$</td>
<td>semi metric</td>
</tr>
<tr>
<td>$s_5$</td>
<td>$\chi/[(\alpha + \beta)(\alpha + \chi)]^{1/\chi}$</td>
<td>$[0,1]$</td>
<td>semi metric</td>
</tr>
</tbody>
</table>

The corresponding dissimilarities are $d_i = 1 - s_i$.

The $d_i$ are aggregated on $p$ variables by the generalised Minkowski metric, thus obtaining:

$$d_i(a, b) = \sqrt[p]{\sum_{j=1}^{p} w_j d_j(A_j, B_j)}$$

$1 \leq i \leq 5$
The object name and the variables names can be moved.
**Method: DIV algorithm**

Divisive and symbolic algorithm

- **THE METHOD**
  - divisive: recursive, descendant
  - binary: binary question
  - symbolic: input: symbolic data
  output: symbolic interpretation
  clusters: assertion object

\[ \sum_{k_i \in C_k} Q(C_k) \]

**Additive criterion**

\[ W(P) = \sum_{C_k \in P} Q(C_k) \]

\( Q \) measures the quality of a cluster

\[ Q(C) = \frac{1}{2n_k} \sum_{a_i \in c_k} \sum_{a_j \in c_k} d^2(\omega_i, \omega_j) \]

\( n_k \) = number of individuals in \( C_k \)

\( d \) = distance or dissimilarity between symbolic objects (Haussdorf, KHI2)

**Normalization:**
- inverse of dispersion (symbolic variance)
- inverse of maximum deviation
Choice of the cut value: $S$

- numerical or ordinal
  - order the value of the variables
  - choice $s$ in the middle of 2 consecutive values
- interval
  - reduce the interval in a point: the center
  - choice $s$: idem numerical method
- probabilistic
  - on probabilistic distribution
  - choice $s = \text{mediane}$

Individual $\omega$ de C:
height($\omega$) = [170, 175]

Output results
Clustering tree
<table>
<thead>
<tr>
<th>PRODUCTS</th>
<th>WEIGHT</th>
<th>COST</th>
<th>AGE</th>
<th>PROFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRODUCT1</td>
<td>[2,4]</td>
<td>[3,5]</td>
<td>[4,6]</td>
<td>[0,3]</td>
</tr>
<tr>
<td>PRODUCT2</td>
<td>[4,5]</td>
<td>[3,4]</td>
<td>[1,6]</td>
<td>[2,7]</td>
</tr>
<tr>
<td>PRODUCT3</td>
<td>[1,6]</td>
<td>[2,7]</td>
<td>[5,8]</td>
<td>[6,9]</td>
</tr>
</tbody>
</table>
Symbolic Principal Component Analysis

Symbolic correlation

[Diagram showing symbolic principal component analysis and symbolic correlation]
SYMBOLIC DESCRIPTION OF CLASSES

- Discrimination
- Classification rules

Decisional *Symbolic* Data Analysis

- Discrimination
- Classification rules

Factorial Discriminant Analysis
(Lauro, Verde, Palumbo, 2000)

Geometrical Visualisation of Symbolic description well separated from each other on a factorial discr plane.

Based on the proximity between objects on factorial planes

Geometrical rules:
- Minimum SO class volume increasing
- De Carvalho’s dissimilarity measure
THE SODAS 2 SOFTWARE FROM ASSO

NEW PROBLEMS APPEAR
- QUALITY, ROBUSTNESS RELIABILITY OF THE APPROXIMATION OF A CONCEPT BY A SYMBOLIC OBJECT,
- THE SYMBOLIC DESCRIPTION OF A CLASS,
- THE CONSENSUS BETWEEN SYMBOLIC DESCRIPTIONS ETC..

MUCH HAS TO BE DONE:
- SYMBOLIC
- REGRESSION, FACTORIAL ANAL.
- MULTIDIMENSIONAL SCALING,
- MIXTURE DECOMPOSITION,
- NEURAL NETWORK, KOHONEN MAP
- CONCEPT PROPAGATION…….
Some recent advances:
- Mixture decomposition of Distributions of distributions (by Copulas, Dirichlet and Kraft stochastic process)
- Stochastic Symbolic Conceptual lattices using capacity theory
- Symbolic class description
- Symbolic Regression
- NEXT FUTUR
  - Spatial symbolic clustering by pyramids
- Symbolic time series.
- Consensus between different description of the same set of units

AIM ATTAINED
FROM HUDGE DATA BASES IN AN ECONOMIC WAY
WE ARE ABLE TO: - Extract new knowledge
  - Summarize
  - Concatenate
  - Solve confidentiality
  - Explain Correlation
HOW? By working on HIGHER LEVEL UNITS extending Data Mining to Knowledge Mining.
CONCLUSION

Symbolic Data Analysis is an extension of standard data analysis therefore
First principle: any Symbolic Data Mining method must have as a special case method of Data Mining on standard data.
Second principle: the output must be a symbolic description or symbolic object
New problems appear as the quality, robustness and reliability of the approximation of a concept by a symbolic object, the symbolic description of a class, the consensus between symbolic descriptions etc..
Due to the intensive development of the information technology the great chapters of the standard statistics will have to be think in these new terms.

References

SPRINGER, 2000 :
“Analysis of Symbolic Data”

JASA (Journal of the American Statistical Association)
“From the Statistic of Data to the Statistic of Knowledge: Symbolic Data Analysis” Billard, Diday June, 2003 .

Electronic Journal of S. D. A.: ESDA
E. Diday, R. Verde, Y. Lechevallier

Download SODAS and SODAS information :
www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm
FROM FUZZY DATA TO SYMBOLIC DATA

<table>
<thead>
<tr>
<th>height</th>
<th>weight</th>
<th>hair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul</td>
<td>1.60</td>
<td>45</td>
</tr>
<tr>
<td>Jef</td>
<td>1.85</td>
<td>80</td>
</tr>
<tr>
<td>Jim</td>
<td>0.65</td>
<td>30</td>
</tr>
<tr>
<td>Bill</td>
<td>1.95</td>
<td>90</td>
</tr>
</tbody>
</table>

Initial Data

<table>
<thead>
<tr>
<th>small</th>
<th>average</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul</td>
<td>0.70</td>
<td>0</td>
</tr>
<tr>
<td>Jef</td>
<td>0</td>
<td>0.50</td>
</tr>
<tr>
<td>Jim</td>
<td>0.50</td>
<td>0</td>
</tr>
<tr>
<td>Bill</td>
<td>0</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Fuzzy Data

<table>
<thead>
<tr>
<th>small</th>
<th>average</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul, Jef</td>
<td>[0, 0.70]</td>
<td>[0.30, 0.50]</td>
</tr>
<tr>
<td>Jim, Bill</td>
<td>[0, 0.50]</td>
<td>[0, 0.48]</td>
</tr>
</tbody>
</table>

From Numerical to Fuzzy Data

Symbolic Data