Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software

Mireille Gettler-Summa[1]

Lise ceremade Université Paris IX Dauphine
1, Place du Ml De Lattre de Tassigny75016 Paris France
summa@ceremade.dauphine.fr

Abstract. In this paper we propose an automatic method of generating Symbolic Objects in the following framework: description of a partition by symbolic objects that takes into account two aspects, that may be called homogeneity and discrimination criteria. This method belongs to a family of algorithms named MGS (Marking and Generalization by Symbolic Objects) in [GETSUM98], which may be applied either to Factorial Analysis interpretation in [GETSUM92][VERGIOGETSUM97], to interpretation of partitions [GETSUMPERFER94], or to summarizing huge databases in

1 MGS in SODAS

1.1 The input classical data matrix

n observations on p nominal variables We consider a set $\Omega = \{1, ...n\}$ of n objects observed for p classical nominal variables Y_j

$$\begin{aligned} Y_j \ &: \ \Omega \rightarrow \mathcal{Y}_j \\ &i \rightarrow Y_j(i) = x_{ij} \end{aligned}$$

$\mathcal{Y}_j$ is the domain of the variable Y_j, that is a finite set of categories
example : Y_j = colour
$x_{ij} \in$ { none, blue, red, yellow, pink}

n observations on p quantitative variables For the symbolic object extraction method which is developed in this paper, only nominal variables are accepted. Quantitative variables should be transformed into nominal ones. If no expert bounds for the intervals exist,an automatic coding is necessary. As all Marking approaches take into account a discrimination criterion (see section and section), the quality of the Markings depends on the chosen bounds for the intervals which will define the categories of the new nominal variables.

We can consider in this quantitative case that we have n data points (vectors) $x_1, ..., x_n \in I\!R^p$ in p-dimensional Euclidean space $I\!R^p$

Let $I\!R^{p-1}$be the $(p-1)$ dimensional Euclidean space containing all the variables, excepting Y_j, and let $\| \quad \|_{p-1}$ denote the corresponding norm.

Let x_i^* be the reduced vector associated with x_i in $I\!R^{p-1}$

Let Y_j be the quantitative variable to be transformed

Let $\{ I_{jk}, 1 \leq k \leq K\}$ be a set of intervals determined for $Y_j$by the coding process.

Let C_{jk} be the extension in Ω of I_{jk} :

$$C_{jk} = \{ i \in \Omega, x_{ij} \in I_{jk}\} \tag{1}$$

We are looking for intervals I_{jk}, $k \in \{1, ...K\}$ such that

$$\mathcal{I}_j = \sum_{k=1}^{K} \sum_{x_k^* \in C_{jk}} \| x_k^* - \overline{x_{c_{jk}}^*} \|_{p-1} \rightarrow \min_{\{I_{jk}\}} \tag{2}$$

Generalized Fisher algorithm (1958) provides a solution to the problem of finding optimized intervals for Y_j . In fact by Fisher process, the extensions of the final intervals have a minimal Within-class Variance with respect to all the remaining variables (and consequently a maximal Between-class Variance with respect to all the remaining variables).

The output of this process is a categorical single-value matrix.

The partition of the initial data

Partition resulting from an automatic clustering process Let assume that $\Omega = \{1, ...n\}$is partitioned into r known disjoint classes $C_1, ..., C_r$, resulting from an automatic clustering process. Each object $k \in \Omega$ of the population is described by

- the p categories of the pattern matrix
- C, the class variable

The variable C is a categorical variable with r levels $\{l_1, ..., l_r\}$.

Externally provided partition Let assume that $\Omega = \{1, ...n\}$ is partitioned into r known disjoint classes $C_1, ..., C_r$, resulting from an external specification, without any reference to an automatic clustering procedure, such as a randomly generated partition, the partition provided by an expert, or the level-induced partition from a categorical variable.

Let C be the partition variable.

Example 1.

- l_1 strong agreement $\rightarrow C_2 = \{\omega_i \in \Omega, C(\omega_i) = l_2\}$
- l_2 mild agreement $\rightarrow C_3 = \{\omega_i \in \Omega, C(\omega_i) = l_3\}$
- l_3 indifference $\rightarrow C_4 = \{\omega_i \in \Omega, C(\omega_i) = l_4\}$
- l_4 mild agreement $\rightarrow C_5 = \{\omega_i \in \Omega, C(\omega_i) = l_5\}$
- l_5 strong agreement $\rightarrow C_6 = \{\omega_i \in \Omega, C(\omega_i) = l_6\}$
- l_6 very strong disagreement $\rightarrow C_7 = \{\omega_i \in \Omega, C(\omega_i) = l_7\}$

In this case, classes are likely to be less homogeneous or/and less isolated than those provided by a clustering process.

As the quality of the discrimination of the Markings depends on the extent to which the classes are separated, it is necessary to look for optimized levels of the partition variable, as an appropriate aggregation of the initial levels in the following sense:

let $L = \{l_1, ...l_r\}$ be the initial set of levels

let $\mathcal{P}'(L)$ be the set of possible partitions of L with at least 2 classes

let $P'_s(L)$ be one of those partitions

$$\begin{cases} \mathcal{P}'(L) \in \mathcal{P}(L) \\ k \equiv Card\left[P'_s(L),\ P'_s(L) \in \mathcal{P}'(L)\right] \\ 2 \leq k \leq l \end{cases} \tag{3}$$

Let C_j^s be the class of Ω induced by the jthclass of $P'_s(L)$

One is looking for a partition $P'_s(L)$ such that the Within-class Variance, according to a chosen metric is minimum:

$$\sum_{1 \leq j \leq k} Var\left(C_j^s\right) \rightarrow \min_s \tag{4}$$

This search can be split into two different situations:

- C is an ordinal categorical variable
- C is an non-ordered (nominal) variable

In both of these cases, experts may sometimes provide a taxonomy of the initial levels.

Example 2. ordered variable (see figure 1.1)

Example 3. non-ordered variable (see figure 1.2)

That is to say that C is a tree-structured variable.

The search for the best partition of the initial levels must consequently be carried out with the constraint of the knowledge of the taxonomy. The search is thus shortened because only partitions which can be derived from the taxonomy are to be examined for the Within class Variance optimization.

A generic procedure may consist of the following steps:

- extract all the partitions of levels available from the given taxonomy
- compute all sums of Within class Variances for each partition
- for the result, choose a partition of levels for which the Within class Variance is minimum

1.2 Marking and Generalization by Symbolic descriptions (MGS) algorithm in SODAS

Let assume C_1 is the class to be marked

Let $\left\{m_j^r, 1 \leq r \leq k_j\right\}$ be the levels of category Y_j

Marking cores are generally formalized by multi-valued Boolean categorical symbolic descriptions which do not necessarily contain the same variables. For SODAS, solely the restricted situation of single valued Boolean descriptions is developed.

Let P_1 be the set of parts of $\mathcal{P}(\{1, ..., p\})$ including solely singletons.

Let denote M_g a generic marking core for C_1, as follows:

$$\begin{cases} M_g \equiv \bigwedge\limits_{j \in P_1} \left[Y_j \in \left\{m_j^{r_g}(1.0)\right\}\right] \\ 1 \leq r_g \leq k_j \end{cases} \tag{5}$$

$\mathcal{Y} = \{$height, sex, grade obtained in an examination, nationality$\}$

levels of height : low, medium, high

levels of sex,: male, female

levels of grade obtained in an examination: A, B, C, D, E

levels of nationality: British, French, German, Italian

$M_1 = [$level of grade obtained in an exmination $\in \{\mathrm{B}(1.0)\}]$

$\bigwedge [\text{nationality} \in \{\text{French}(1.0)\}]$

$M_2 = [\text{height} \in \{\text{high}(1.0)\}]$

$M_3 = [\text{height} \in \{\text{low}(1.0)\}]$

$\bigwedge [\text{nationality} \in \{\text{French}(1.0)\}] \bigwedge [\text{sex} \in \{\text{male}(1.0)\}]$

Two categories are missing in M_1, three in M_2, one in M_3

Variables values which are not specified are neither missing values, nor values having all possible values in their domain. In the framework of SODAS, Symbolic Object Language requires that marking cores were completed on the not specified variables.

See section 1.2.6 for the details on the option that has been chosen in SODAS.

Results are thus formalized as modal multivalued symbolic descriptions

Three criteria for constructing the marking cores One is looking for conjunctions of initial levels (each conjunction will be a marking core) such that:

(i) the cardinal of the union of their extensions in C_1 is maximal

(ii) the cardinal of the union of their extension in $C_{_\Omega}(C_1)$

(iii) each hypercube representing conjunctions of levels is statistically significant with respect to the test which has been chosen for measuring the quality of the linkage between C_1 and markings.

C_1class constituent objects are labelled 1

Three markings have been constructed on C_1

According to A. Gordon,[GOR99] presentation, (i) and (ii) can be written as follows:

(i) Minimize false negatives, i.e. objects belonging to the extension of a marking but not to C_1

(ii) Minimize false positives, i.e. objects belonging to C_1 but not to the union of the extensions of markings

In figure there are on the whole:

- four false negatives
- two false positives, one in M_1, one in M_2; M_3 has no false positive

For the third criterion (iii), various measures can be computed for measuring the link between a generic Marking M_g and C_1.

Almost all of them use the following quantities:

$n_g = Card\left[ext_\Omega\left(M_g\right)\right], n_{1g} = Card\left[ext_{C_1}\left(M_g\right)\right], n_{g-}n_{1g} = Card\left[ext_{C_\Omega C_1}\left(M_g\right)\right]$

	C_1	$C_\Omega(C_1)$	population Ω
M_g	n_{1g}	$n_{g-}n_{1g}$	n_g
population Ω	n_1	n-n_1	n

(6)

- the χ^2 of contingency is a measure of the deviation from independence between M_g and C_1

$$\chi^2_{1,g} = \frac{\left(n_{1g} - n\frac{n_g}{n}\frac{n_1}{n}\right)^2}{n\frac{n_g}{n}\frac{n_1}{n}} \tag{7}$$

the nearer to zero $\chi^2_{1,g}$ is, the weaker the link is; the theoretical expected value, under the independence null hypothesis is equal to $\frac{n_g n_1}{n}$

- the Test-Value of Morineau [ALMOR92] is based on hypergeometric; for its 5% upper point, its value, denoted T_H, is superior or equal to 1.96; a Laplace Gauss approximation is given by:

$$T_H = \frac{n_{1g} - n\frac{n_g}{n}}{\left[n_1\frac{n-n_1}{n-1}\frac{n_g}{n}/1 - \frac{n_g}{n}\right]^{1/2}} \tag{8}$$

This formula cannot be used when occurrences are too few

Those quantities are used for ranking the Markings with respect to the quality of the link they have with the class to be marked. The greater they are, the more relevant are the corresponding markings to the class.

The algorithmic approach Various heuristics have already been proposed to construct Marking cores. Main differences are wether they are top down [BLANCGETSUM94]or bottom up [PHAMGETSUM98], greedy [HODIDSGETSUM88] or not, depth first or breadth first, allowing overlapping branches or not etc.

Let denote L=$\left\{\text{lg}, 1 \leq g \leq v\right\}$the set of the levels of all the variables.

Let denote S_M a set of Markings.

Let denote $Cov(l_g) \equiv Card\left[ext_{C_1}\left(l_g\right)\right]$(first criterium)

Let denote $Err(l_g) = Card\left[ext_{C_\Omega\{C_1\}}(l_g)\right]$(second criterium)

Two a priori thresholds are to be chosen:

- the final degree in which C_1 is covered by the union of the markings
- the errors made by the markings by covering elements out of C_1

Let denote r_{Cov} the first threshold; the ratio for covering obtained by the final markings should be such that:

$$\frac{ext\left[\bigcup_g M_g\right]}{Card(C_1)} \geq r_{Cov} \tag{9}$$

Let denote r_{Err} the second threshold; error ratio for a marking should be such that:

$$\forall M_g \in S_M \qquad \frac{Err(M_g)}{ext_\Omega(M_g)} \leq r_{Err} \tag{10}$$

STEP 1

- levels are ordered by their measures in the framework of criterium (iii). Let denote $T\left(l_g\right)$ this value for a generic level l_g
- all first levels build a first set of marking cores, which will eventually be improved at further steps; l_g can thus be denoted as M_g^1. Criteria (i) and (ii) are computed for each marking:

$$Cov(M_g^1) \equiv Card\left[ext_{C_1}\left(M_g^1\right)\right] \tag{11}$$

$$Err(M_g^1) = Card\left[ext_{C_\Omega(C_1)}(M_g^1)\right] \tag{12}$$

If a level is such that $Err(l_g)/Card(\Omega) \geq r_{Err}$, the corresponding marking is abandoned for the further steps.

A first set of marking cores is thus constructed:

$$\begin{cases} S_M^1 = \left\{M_g^1, M_g^1 \equiv l_g, 1 \leq g \leq v_1 \leq v\right\} \\ Card(S_M^1) = v_1 \\ \left(Err(M_g^1)/Card(\Omega)\right) \leq r_{Err} \end{cases} \tag{13}$$

The two following quantities are also computed:

$$Cov\left(S_M^1\right) \equiv Card\left\{ext_{C_1}\left(\cup_{S_M^1} M_g^1\right)\right\} \tag{14}$$

$$Err\left(S_M^1\right) \equiv Card\left\{ext_{C_\Omega(C_1)}\left(\cup_{S_M^1} M_g^1\right)\right\} \tag{15}$$

STEP 2

– Each element of S_M^1 will be a root for descending branches built as follows.
– The constituents of S_M^1 are ordered by their corresponding values

$\left\{T\left(M_g^1\right), 1 \leq g \leq v_1\right\}$(third criterium)

$$T\left(M_{g_1}^1\right) \geq T\left(M_{g_2}^1\right) \geq ... \geq T\left(M_{gv_1}^1\right) \tag{16}$$

– The greatest value corresponds to the root which is processed at first and so on.
– Branches are constructed from each node by choosing the levels with the above defined order (see figure 1.3)
– For each branch, one has to check if it has not yet been constructed for avoiding redundancy (see figure 1.4)
– For each branch, the error ratio is computed; if it is greater than a priori threshold, the branch is abandoned
– Each branch as a wholeis a new marking

A second set of marking cores is thus substituted to the first one:

$$\begin{cases} S_M^2 = \left\{M_g^2, M_g^2 \equiv l_g, 1 \leq g \leq v_2\right\} \\ Card(S_M^2) = v_2 \\ \left(Err(M_g^2)/Card(\Omega)\right) \leq r_{Err} \end{cases} \tag{17}$$

The following quantities are also computed:
for each marking M_g^2 $(1 \leq g \leq v_2)$: $T\left(M_g^2\right)$ (criterium (iii))
for S_M^2,

$$Cov\left(S_M^2\right) \equiv Card\left\{ext_{C_1}\left(\cup_{S_M^2} M_g^2\right)\right\} \tag{18}$$

$$Err\left(S_M^2\right) \equiv Card\left\{ext_{\Omega(C_2)}\left(\cup_{S_M^2} M_g^2\right)\right\} \tag{19}$$

FURTHER STEPS

Step 2 procedure is iterated and stops according to the stopping rules which are described in the following paragraph.

Stopping and non stopping rules As the number of starting levels is limited and redundancy of branches is avoided, the algorithm naturally proceeds with a finite number of steps and gets to an end.

Some stopping rules can shorten the process:

– a step f is the last one if $Cov\left(S_M^f\right)/Card_{C_1}\left(S_M^f\right) \geq r_{Cov}$

i.e. C_1 has been sufficiently marked

– if one does not want more than h levels in a description (for example for providing a quick decision aid rule in an application) no branches will be developed after the h^{th} step which is at the most the last one
– if a final marking M_f is such that $Err\left(M_f\right)/Card_{C_1}\left(M_f\right) \geq r_{Err}$, it can be cancelled, as an option of the algorithm, from the results

Marking cores The markings which are the results of the above process are the so called marking cores for the class C_1.They are Boolean descriptions, such that each mentioned level has a hundred per cent presence in the description.

Example 4. One is looking for summarizing by their answers the class of respondents, who agree with the politics of management of their city

respondents	education level	opinion on administrator personality	opinion on city administration	sex	partition
1	A	good	good	female	agreement
2	B	good	good	female	agreement
3	A	medium	good	female	agreement
4	C	good	indifferent	male	agreement
5	D	good	good	male	agreement
6	C	good	bad	male	agreement
7	C	bad	good	male	disagreement
8	B	good	bad	male	disagreement
9	A	bad	bad	male	disagreement
10	C	bad	good	female	disagreement
11	A	good	bad	female	disagreement

(20)

There are eleven respondents, who answer to five categorical questions:
levels of variable labelled "education level": A, B, C, D
levels of variable labelled "opinion on administrator personality": good, medium, bad
levels of variable labelled "opinion on city administration": good, indifferent, bad
levels of variable labelled "sex": male, female
levels of variable labelled "partition": agreement, disagreement
Construction of the marking cores for the class C_1 "agreement":
ordering levels, simply according to their frequencies:

opinion on administrator personality	good	5
opinion on city administration	good	4
sex	male	3
sex	female	3
education level	A	2
education level	C	2
etc.		

(21)

Some first branches, top down and left first (see figur 1.5)
The final markings depend on the values of r_{Cov} and r_{Err}

Remark 1. If $r_{Err} = 1.0$, then markings may make no discrimination at all between C_1 and the other classes, and the markings are thus simply generalizations of C_1.

For $r_{Cov} = 80\%$ and $= 25\%$,the set of marking results is:

$$\begin{cases} M_1 = & [\text{opinion on administrator personality} \in \{\text{good}\,(1.0)\}] \\ & \wedge [\text{opinion on city administration } \in \{\text{good}\,(1.0)\}] \\ M_2 = & [\text{opinion on administrator personality} \in \{\text{good}\,(1.0)\}] \\ & \wedge [\text{sex} \in \{\text{male}\,(1.0)\}] \\ M_3 = & [\text{opinion on city administration} \in \{\text{good}\,(1.0)\}] \\ & \wedge [\text{sex} \in \{\text{female}\,(1.0)\}] \end{cases} \tag{22}$$

There are so three marking cores.

$M_1$and M_2 are overlapping: one elements (repondents 5) belongs to both of them

M_3 and $M_1$are overlapping: two elements (repondents 1 and 2)

Almost 83% (respondents 1,2,4,5,6) of the elements of $C_1$are described by the whole set of markings.

There is one false negative (respondent 7), that is hardly 17% error, and one false positive (respondent 4)

Two markings ($M_1$and M_3) make no error, whereas one marking (M_2) has a 25% error ratio.

Unions and intersections One may compute the union of two markings which are widely overlapping or else compute appropriate intersections of markings.The indexes of Covering and Error should then be computed for these new elements because they measure the quality of a marking. If they are satisfactory, one can keep these new elements as results, while crossing out the elements which are their components.

In this situation, different levels of a same variable may be put into disjunction in marking; results are thus multivalued symbolic descriptions.

From marking cores to full specified symbolic description Depending on the way variable values are completed on not specified categories, marking cores are more or less generalised to modal multivalued descriptions.

Actually, the choice which has been implemented in SODAS consists in subsituting a missing value by the (discrete) distribution of the missing category on the extension of a marking core in class C_1

The completed markings M_1^*, M_2^*, M_3^* corresponding to the three markings M_1, M_2, $M_3$of previous example are the following:

$$\left\{\begin{array}{l} M_1 = \quad [\text{opinion on administrator personality} \in \{\text{good}\,(1.0)\}] \\ \qquad \wedge\,[\text{opinion on city administration} \in \{\text{good}\,(1.0)\}] \\ \qquad \wedge\,[\text{education level} \in \{\text{A}\,(0.25)\,,\text{B}\,(0.25)\,,\text{C}(0.25),\text{D}\,(0.25)\}] \\ \qquad \wedge\,[\text{sex} \in \{\text{female}\,(0.5)\,,\text{male}\,(0.5)\}] \\ M_2 = \quad [\text{opinion on administrator personality} \in \{\text{good}\,(1.0)\}] \\ \qquad \wedge\,[\text{sex} \in \{\text{male}\,(1.0)\}] \\ \qquad \wedge\,[\text{education level} \in \{\text{C}(0.67),\text{D}\,(0.33)\}] \\ \wedge\,[\text{opinion on city administration} \in \{\text{good}\,(0.67)\,,\text{indifferent}\,(0.33)\}] \\ M_3 = \quad [\text{opinion on city administration} \in \{\text{good}\,(1.0)\}] \\ \qquad \wedge\,[\text{sex} \in \{\text{female}\,(1.0)\}] \\ \qquad \wedge\,[\text{education level} \in \{\text{A}(0.67),\text{B}\,(0.33)\}] \\ \wedge\,[\text{opinion on administrator personality} \in \{\text{good}\,(0.67)\,,\ \text{medium}(0.33)\}] \end{array}\right. \tag{23}$$

Performance of the algorithm Some work has been carried on the performance of MGS algorithm by comparing different indexes for measuring the quality of the link between markings and class C_1 (Pham Ti Tong, Gettler-Summa, 1996)

Experimentation has been done on four different data sets coming from the UCI Machine Learning Repository site (ftp://ics.uci.edu/pub): WINE, VOTE, WAVE, ZOO. Data have been processed with four different indexes for criterium (iii): the Test-value based on the hypergeometric statistic, the Shannon entropy, the J-measure, and the χ^2.

Each data set Ω is randomly divided into 10 mutually exclusive subsets Ω, Ω_2,..., Ω_{10} of approximatively equal size. Each measure is tested 10 times. At each time, marking cores are constructed on Ω
Ω_k.

The following table presents the average of the error rate of miss classifications:

	Test-value	Shannon entropy	J-measure	χ^2
WINE	0.05	0.08	0.08	0.09
VOTE	0.06	0.08	0.6	0.14
WAVE	0.26	0.18	0.20	0.35
ZOO	0.50	0.47	0.23	0.40

(24)

1.3 Official Statistical Institute And Industrial Applications

Official Statistical Institute applications An extract of INE (National Istitute of Statistic of Portugal) Labor Force Survey have been processed: 2193 units and 34 categories such that

look for job,3 levels
fullpart,9 levels
principal activity, 12 levels

+months,6 levels
sex,3 levels
etc.

An external partition has been given, in two classes: employed, unemployed.

Here is an example of a Marking core which has been obtained as a result through MGS procedure:

princact$\in \{pers_serv\&sic\}$)
"and"(prinprof$\in \{pers_serv\&sic\}$)
"and"(fullpart$\in \{full\}$)
"and"(act$\in \{yes\}$)
"and"(inscr$\in \{no\}$)
"and"(lookforjob$\in \{no\}$)
"and"(bestway$\in \{nr\}$)
"and"(months$\in \{NA\}$)
"and"(typesec$\in \{NA\}$)

This Marking covers 9.6% of the respondents of the class "employed", with no error at all; its Test-Value is equal to 16.91

The associated completed markings have then be processed as inputs for a Symbolic Discriminant Factorial Analysis through SODAS

Industrial applications Many industrial applications and different implementation of the method in softwares have already been realised [BLANCGETSUM94][GETSUMVAUT97] [GETSUM97][MORGETSUMPHAM96]

The results we are here describing come from NOEMIE european contract on industrial feedback process. A Data Warehouse is buit to provide final tables on which data mining is held to discover hidden regularities.

Let present the Markings of the finantial relational data base of around 40 000 units and 12 attributes which keeps the memory of the sells of tools in the company for at least ten years: 265 markings have been constructed from 80 classes provided by automatic clustering.

Here is an example of Marking core on one class E (Step 1 of the algorithm):

(Customer$\in \{SKA\}$)
"and"(Price$\in \{[0, 100]\}$)
"and"(quantity$\in \{[50, 100]\}$)

Some Markings related to the same class E are then grouped according to the T_H test. Richer Markings are obtained on E, which are descriptions of multivalued symbolic Boolean objects :

(Customer$\in \{SKA, DIE\}$)
"and"(Price$\in \{[0, 100], [100, 1000]\}$)
"and"(quantity$\in \{[50, 100]\}$)

In this application, experts have asked to complete Markings on some variables (not all as requested by the Symbolic Object Linguage of SODAS)which could have been irrelevant for the Markings (branch stopped from a statistical point of view, before this variable appears), but important from an expert point of view. Here is a new Marking corresponding to previous one:

(Customer$\in \{SKA, DIE\}$)
"and"(Price$\in \{[0,100],[100,1000]\}$)
"and"(quantity$\in \{[50,100]\}$)
"and"(Tool$\in \{AAKB, AFAC\}$)
"and"(Year$\in \{1995, 1996\}$)

The next step of the algorithm calculates the frequency distribution of each category involved in the markings on the basis of the extension of the marking on the class E or on the set Ω :

(Customer$\in \{SKA(0.7), DIE(0.3)\}$)
"and"(Price$\in \{[0,100](0.9),[100,1000](0.1)\}$)
"and"(quantity$\in \{[50,100](1.0)\}$)
"and"(Tool$\in \{AAKB(0.8), AFAC(0.2)\}$)
"and"(Year$\in \{1995(0.5), 1996(0.5)\}$)

The Markings are then used as querries to summarize one state of the data base.

References

[ALMOR92] P. ALEVIZOS, A. MORINEAU, "Tests et Valeurs-Tests", RSA vol 40, 1992

[BLANCGETSUM94] J.L. BLANCHARD, M. GETTLER-SUMMA, "Symbolic approaches on ergonomic problems for electric centres", Compstat, Vienna, July 1994.

[DIDAY95] , "Probabilistic objects for a symbolic data analysis", Series in Discrete Mathematics and Theoretical Computers, 19, 1995.

[HODIDSGETSUM88] B. Ho Tu, E. DIDAY, M. GETTLER-SUMMA "Generating rules for expert systems from observations" Pattern Recognition Letters n°7, 1988

[GETSUM92] M. GETTLER-SUMMA, "Factorial axis interpretation by symbolic objects" III^{èmes} Journées numérique-symbolique, Université Paris IX-Dauphine, CEREMADE, (Ed), 1992.

[GETSUMPERFER94] M. GETTLER-SUMMA, E.PERINEL,J.FERRARIS, "New automatic aid to symbolic cluster interpretation",New Approaches in Classification and Data Analysis, Springer Ed., 1994

[GETSUM97] M. GETTLER-SUMMA, "Symbolic Marking ; Application on car accident scenarios". Proceedings of Applied Stochastic Models and Data Analysis, Capri, Italy, 1997.

[GETSUM98] M. GETTLER-SUMMA, "Approches MGS : Marquage et Généralisation Symbolique pour de nouvelles aides à l'interprétation en Analyse de Données", Cahiers du Ceremade N^o 9830, Université Paris IX-Dauphine, France, 1998.

[GETSUMVAUT97] M. GETTLER-SUMMA, F. VAUTRAIN,"Discrimination d'exemples très rares dans la base d'apprentissage- SFC Vannes 1997 France

[GOR99] A.D.GORDON, "Classification (2nd Edition), Chapman&Hall/CRC,Boca Raton,FL. 256pp.,,1999

[PHAMGETSUM96] H.PHAM TI TONG, M. GETTLER-SUMMA, " a bottom up procedure for generating rules", IFCS 96, Kobe,Japan, 1996

[PHAMGETSUM98] H.PHAM TI TONG, M. GETTLER-SUMMA, "Performances d'un algorithme générateur de règles", SFC 98, France, 1998

[MASGETSUMDIDTOUAT98] M.MASSRALI, M. GETTLER-SUMMA, E.DIDAY, M. TOUATI, "extracting knowledge from very large data base",KESDA, Luxembourg 1998

[VERGIOGETSUM97] R. VERDE,F.GIORDANO,M. GETTLER-SUMMA, "Symbolic objects for multiattribute preference data", OSDA 97, Darmstadt (Germany), 1997.

[MORGETSUMPHAM96] A. MORINEAU, M.

GETTLER-SUMMA, H. TONG, "Marquage sémantique des classes et des axes", XXVII^{èmes} Journées de l'ASU, Paris, 1996.