

Generalized Symbolic Marking Of Complex Objects Through Intelligent Complex Miner Software CRM Applications

Mireille Gettler Summa¹, Frederick Vautrain², and Matthieu Barrault²

¹ Centre de Recherche de Mathematique de la Decision
University of Paris Dauphine
1 Pl. du Ml. De Lattre de Tassigny
75016 - Paris France
(e-mail : summa@ceremade.dauphine.fr)

² ISTHMA Ltd
6 rue du Soleillet
75020 - Paris France
(e-mail: barrault@isthma.com, vautrain@isthma.com)

Abstract. In this paper we propose an automatic method of describing classes of complex objects (lists, diagrams, intervals, histograms, time series). The approach simultaneously generalizes a class and discriminates it from the others. This method belongs to a family of algorithms called MGS (Marking and Generalization by Symbolic objects) which were already applied on classical inputs, either to Factorial Analysis interpretation in [Gettler Summa, 1992] [Giordano *et al.*, 2000] or to the interpretation of partitions [Gettler Summa *et al.*, 1994]. It was also used for summarizing huge databases in [Massrali *et al.*, 1998]. For Customer Relationship Management, MGS provides sets of client profiles which target shops, brands or couponing analysis. An application through Intelligent Complex Miner software is presented on jointed data bases of sells, couponing information, client socio-demographic elements, and geo-marketing data.

Keywords: discriminant description, generalization, symbolic marking, generalized V-test, CRM.

1 Introduction

Data analysis on classes of statistical units is a crucial issue because huge data bases are stored and results interpretation takes more and more time. Furthermore, to take into account multiple arrays, distribution values, time series or continuous functions appear to be the very appropriate inputs for summarizing the data without loss of information, towards knowledge extraction. Very few approaches face such complex data analysis: Symbolic Data Analysis brings for example a theoretical framework [Diday, 1988] for such a challenge. Discrimination and generalisation are to be redefined in that new context: Marking and Generalisation by Symbolic objects generalises to symbolic inputs some Machine Learning approaches [Stepp, 1984] [Ho tu *et al.*, 1988] [Ganascia, 2000], or supervised classification algorithms that are

generally not able to treat complex matrices [Gordon, 1999]. MGS provides discriminant generalizing objects that describe subsets of the power set of an initial classical data set. Some other symbolic approaches have recently been published [Vrac and Diday, 2001] for similar purposes. As in recursive partition algorithms [Périnel *et al.*, 2003], the results could be written as complex production rules but the inference validation phase is not included in this paper.

2 The input matrix

We consider a set $\Omega = \{\omega_1, \dots, \omega_n\}$ of n symbolic objects for p variables Y_j :

$$\begin{aligned} Y_j &: \Omega \longrightarrow \mathcal{Y}_j \\ \omega &\longrightarrow Y_j(\omega) \end{aligned}$$

$Y_j(i)$ may be :

- a single real number or a single category (classical data);
- a finite set of real numbers or categories (multi-valued variable);
- a discrete finite frequency distribution (diagram variable); frequencies are in this paper the frequencies of a statistical distribution, as it happens for example when symbolic objects come from a query on a classical categorical data base;
- an interval;
- a continuous frequency distribution on a finite number of intervals (histogram variable); an hypothesis of uniform distribution along the intervals allows linear interpolation to calculate frequencies on sub intervals.

Let E and \bar{E} be two classes of a binary partition on Ω (if the given partition is not binary, E is the class to be marked and \bar{E} its complementary part, union of the other classes of the partition).

$Y(\omega) = \{y_j(\omega), j \in \{1, \dots, p\}\}$ is the description of ω , denoted also d_ω . A 'partial description' has fewer variables than in initial individuals. A symbolic object s , is a triplet (a, R, d) where a is a mapping $E \rightarrow \{0, 1\}$ which measures the fit between d_ω and d , R is a relation which associates (in this paper) to a couple of descriptions a Boolean value (for example R is the inclusion operator). The extent of a symbolic object s in E is defined in this paper in the Boolean case by : $Ext_E(s) = \{\omega \in E / a(\omega) = 1\}$.

Let S_E be the set of symbolic objects belonging to E .

3 GV-TEST criterion

Marking is a process which builds discriminant descriptions of S_E . Several trees are simultaneously explored top down from initial nodes. Depending on some parameter values, final descriptions:

1. may be totally discriminant or only partially;
2. may have overlapping extents or not;
3. may include in their extent all E elements or not.

The first and the second points are simultaneously taken into account in the descending process through a threshold Thr_{Cr} for some criterion Cr which measures the link between any subset of Ω and E .

Let M_g be an intent of a subset in the case of a classical initial data set (for example: 'colour = yellow and length = short', partial description for all 'yellow and short' units).

Almost all of them use for classical data the following quantities:

$$n_g = Card[ext_{\Omega}(M_g)], n_{E,g} = Card[ext_E(M_g)], n_g - n_{E,g} = Card[ext_{\bar{E}}(M_g)]$$

	E	\bar{E}	Ω
M_g	$n_{E,g}$	$n_g - n_{E,g}$	n_g
Ω	n_E	$n - n_E$	n

Table 1. GV-TEST criterion

We propose the GV-test criterion which is a generalization of the V-test [Alevizos and Morineau, 1992] for symbolic objects.

The V-test is based on the hyper geometric distribution hypothesis; for its 5% upper point, its value, which can be calculated by a Laplace Gauss approximation, is greater or equal to 1.96. Explicit formula is the following :

$$T_H = \frac{n_{E,g} - n \frac{n_g}{n_E}}{\left[\frac{n_E n_g (n - n_E)}{(n - 1)n \left(1 - \frac{n_g}{n}\right)} \right]^{\frac{1}{2}}}$$

In the case of categorical variables, the V-test criterion can be used on not too small data sets [Gettler Summa, 2000]. For non classical objects, some other calculations (extents cardinalities) are to be considered depending on the situation:

- for an interval variable Y , the frequency of an interval I_1 among all possible intervals for Y values (in E , in \bar{E} , in Ω , in M_g) is equal to the number of intervals that include I ;
- for a multi-valued variable, the frequency of a list L among all possible lists for Y values (in E , in \bar{E} , in Ω , in M_g) is equal to the number of lists which include L ;

- for a diagram variable Y , the frequency which occurs is that of a class of bars of the following type(see 'initial nodes for the sets of diagrams variables'): $\{m_k, w \geq w_{min,m_k}\}$; this cardinality is equal to the number of bars among all Y diagrams (in E , in \bar{E} , in Ω , in M_g) $\{m_{ki}, w_{ki}\}$ where $m_{ki} = m_k$ and $w \geq w_{min,m_k}$;
- for a histogram variable Y , the frequency is the one of a class of rectangles of the following type(see 'initial nodes for the sets of histograms variables') : $\{I_k, w \geq w_{min,I_k}\}$; this cardinality is equal to the number of rectangles among all Y histograms (in E , in \bar{E} , in Ω , in M_g) $\{I_{ki}, w_{ki}\}$ where $I_k \subseteq I_{ki}$ and $(I_k/I_{ki})w_{ki} \geq w_{min,I_k}$.

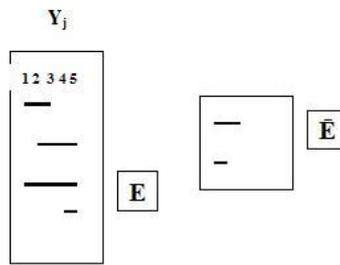


Fig. 1. Example of GV-test computation for an interval variable.

The GV-test of a variable value is used for ranking the initial nodes in order to begin the descending process in the MGS algorithm.

For example, one can easily calculate from figure 1 the elements which occur for the GV-Test computation, for variable Y_j taking its value in the interval $[4, 5]$: Let $s_{[4,5]}$ be the symbolic object associated to the partial description $d = (Y_j = [4, 5])$. $n_g = Card[ext_{\Omega}(s_{[4,5]})] = 3, n_{E,g} = Card[ext_E(s_{[4,5]})] = 3, n_g - n_{E,g} = Card[ext_{\bar{E}}(s_{[4,5]})] = 0, n_E = 4, n = 6$.

4 Initial Nodes

The choice of the initial nodes depends on the variables nature. Details are given for diagram and histogram variables.

4.1 Initial nodes for the sets of categorical or continuous classical variables

Continuous variables are discretized into optimized classes by a supervised method in order to take into account the binary partition: for example the supervised Fisher algorithm or any decision tree one-to-one variable [Zighed *et al.*, 1997]. Initial nodes are then built by the same method as for categorical variables, each class being considered as a category. This approach on

classical data is fully described in [Gettler Summa, 2000]. Let IN_C be the subset of the initial nodes.

4.2 Initial nodes for the set of diagram variables

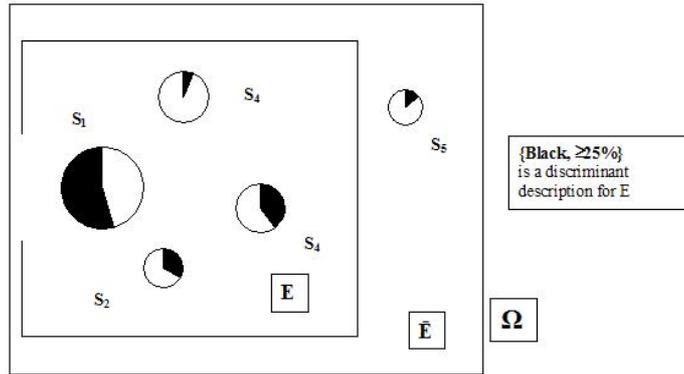


Fig. 2. Initial nodes for the set of diagram variables.

Let $S_{w_c, E}$ be the set of all weighted categories belonging to the variables describing S_E and p_{w_c} be the number of the distinct categories in all diagram variables.

Let's call IN_D the subset of $S_{w_c, E}$ which belong to the initial nodes set.

Let w_1 and w_2 be two weights of a same category, m_k ; suppose $w_1 > w_2$.

Taking into account $\{m_k, w \geq w_2\}$ implies taking into account $\{m_k, w = w_1\}$ because weights have a statistical frequency semantic.

Let then w_{min, m_k} be the minimum of m_k weights that correspond to GV-test values greater than the a priori threshold Thr_{GV} . It may happen for some category that no weight provides a good GV-test. Let p'_{w_c} be the number of those categories m_k for which w_{min, m_k} does exist ($p'_{w_c} \leq p_{w_c}$).

IN_D is then built of all $\{m_k, w \geq w_{min, m_k}\} k \in 1, \dots, p'_{w_c}$.

4.3 Initial nodes for the set of histogram variables

A simple situation is the one of interval variables as it is shown in the GV-test paragraph. It will thus not be detailed furthermore so let us present directly the case of histogram variables.

Let $S_{H, E}$ be the set of all weighted intervals (I_k, w_k) , belonging to S_E .

Let IN_H (IN_I for interval variables) be the subset of $S_{H, E}$ which will belong to the initial nodes set.

Let p_H be the cardinality of $S_{H, E}$.

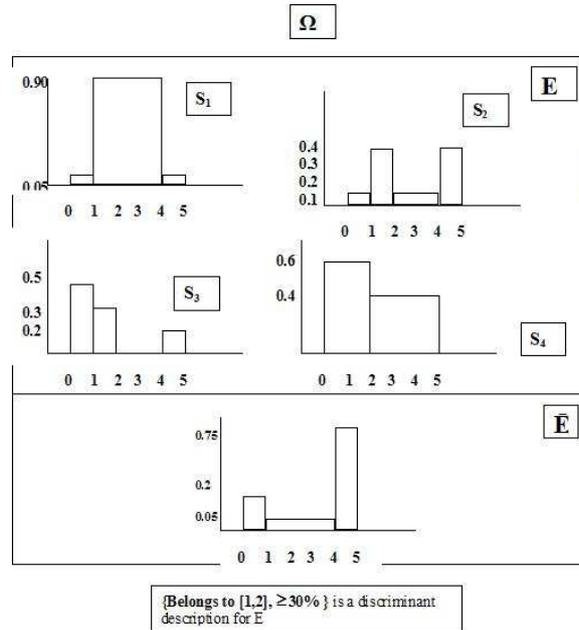


Fig. 3. Initial nodes for the set of histogram variables.

Let us consider the ordered list of the ordered bounds of the intervals $\{(I_k, w_k), k \in 1, \dots, p_H\}$. That list defines $p'_H = p_H$ non overlapping intervals $\{I'_k, k \in 1, \dots, p'_H\}$. Let SI_k be the set of weighted intervals including I'_k . For each of SI_k intervals, a new weight is calculated by linear interpolation (because of the uniform distribution hypothesis). Let w_1 and w_2 be two weights of a same interval, I'_k , depending of two corresponding $S_{H,E}$ intervals the intersection of which is I'_k ; let suppose $w_1 > w_2$. Considering $\{I'_k, w \geq w_2\}$ implies considering $\{I'_k, w = w_1\}$ because weights have a statistical frequency semantic. Let w_{min, I'_k} be the minimum of SI_k weights that correspond to GV-test values greater than the a priori threshold Thr_{GV} . Note that it may happen for some interval that no weight provides a good GV-test. Let p''_H be the number of those intervals for which w_{min, I'_k} does exist ($p''_H \leq p'_H$). IN_H is then built of all $\{(I'_k, w \geq w_{min, I'_k}), k \in 1, \dots, p''_H\}$.

5 Generalized MGS Algorithm

In the proposed method a set of initial nodes is built by first exploring a particular lattice, and then by the conjunction of some vertices, according to various chosen criteria, further steps build the final Markings set.

'The output of the marking process consists of a set of partial descriptions ('Markings'), the number of which is inferior to the number of initial individuals.' Let now call a partial description d and the symbolic object, triplet (a, R, d) that is associated to it by the same notation .

Let's denote M_g a generic marking for E . A threshold Thr_{GV} for the GV-test is to be chosen as an input parameter in order to select the initial nodes for each type of variable. Let us denote L the union of initial nodes sets of different types (see 'initial node'):

$$\begin{aligned} L &= IN_C \cup IN_I \cup IN_D \cup IN_H \\ L &= \{l_g, 1 \leq g \leq v\} \end{aligned}$$

Each of L elements has a GV-test value (see GV-test) with respect to E , such that GV-test values are not metric but ordered values, i.e. the greater the absolute value is, the stronger is the link which is measured.

L elements are thus ordered according to their V-test values.

Various heuristics have already been proposed to construct Markings. The main differences are, whether it is top down or bottom up [Gettler Summa *et al.*, 1995], greedy [Ho tu *et al.*, 1988] or not, depth first or breadth first, allowing overlapping branches or not etc.

Let's denote S_M a set of Makings.

Let's denote $Cov(l_g) = \frac{Card[ext_E(l_g)]}{Card(E)}$.

Let's denote $Err(l_g) = \frac{Card[ext_{\bar{E}}(l_g)]}{Card(\Omega)}$.

Two a priori thresholds are to be chosen:

- The final degree in which E is covered by the union of the markings, R_{Cov} ; a final marking set S_M should be such that:

$$R_{Cov} \leq \frac{Card(\cup ext_E[M_g, M_g \in S_M])}{Card(E)} \quad (1)$$

- The error ratio made by the markings by covering elements out of E , R_{Err} ; each marking should be such that:

$$\forall M_g \in S_M, R_{Err} \geq \frac{Card[ext_{\bar{E}}(M_g)]}{Card(\Omega)} \quad (2)$$

Step 1 : All initial nodes build a first set of markings. Criteria (1) and (2) are calculated for each marking. If any node does not respect Criterion (2), it is deleted from the markings. A first set of markings is thus constructed:

$$S_M^1 = \{M_g^1, M_g^1 = l_g, 1 \leq g \leq v_1 \leq v\}$$

$$Card(S_M^1) = v_1$$

$$\forall M_g^1 \in S_M^1, \frac{Err(M_g^1)}{Card(\Omega)} \leq R_{Err}$$

The two following quantities are also calculated:

$$Cov(S_M^1) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^1)]}{Card(E)}$$

$$Err(S_M^1) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^1)]}{Card(\Omega)}$$

Step 2 : Each element of S1M will be a root for descending branches built as follows :

- the constituents of S1M are ordered by their corresponding GV-test values;
- the greatest GV-test value corresponds to the root which is processed at first and so on;
- branches are constructed from each node by choosing the L elements according to the above defined order;
- for each branch, one has to check if it has not yet been constructed to avoid redundancy;
- for each branch, the error ratio is calculated; if it is greater than RErr, the branch is abandoned;
- for each branch, the GV-test is calculated; if it is smaller than ThrGV, the branch is abandoned;
- each remaining branch as a whole is a new marking.

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

$$S_M^2 = \{M_g^2, 1 \leq g \leq v_2\}$$

$$Card(S_M^2) = v_2$$

$$\forall M_g^2 \in S_M^2, \frac{Err(M_g^2)}{Card(\Omega)} \leq R_{Err}$$

The two following quantities are also calculated:

$$Cov(S_M^2) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^2)]}{Card(E)}$$

$$Err(S_M^2) = \frac{Card[\cup ext_E(M_g, M_g \in S_M^2)]}{Card(\Omega)}$$

Further steps : Step 2 procedure is iterated; as the number of nodes is limited by the GV-test criterion and redundancy of branches is avoided, the algorithm is not fully combinatory and comes to an end according to some stopping rules which are described in the following paragraph:

- a step f is the last one if $Cov(S_M^f) \geq R_{Cov}$ i.e. E is sufficiently marked;

- if one does not want long branches (for example for providing a quick decision aid rule in an application), a parameter is proposed in input to fix a maximum h_{max} for the number of nodes in a branch. The h^{th} step will thus be at the most, the last one;
- if a marking M_f is such that $Err(M_f) \geq R_{Err}$, it can be cancelled, as an option of the algorithm, from the results.

6 Applications

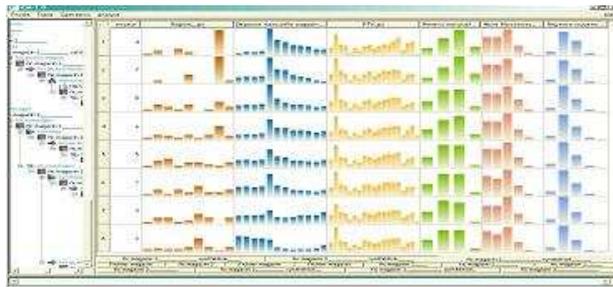


Fig. 4. Complex data editor

MGS is one of the statistical methods implemented in the Interactive Complex Miner software. Based on a collaboration between Ceremade laboratory from Dauphine University and Isthma Company, this software can be used to manage and analyse complex and "multivalued" data as curves, distributions, intervals, sets as well as classical data. Temporal and geographical data are the most frequent complex data types which are analysed in the applications.

Current application consists in three related data bases: 200 shops described by their monthly turnover on several years, 800000 households described by sociogeographic variables and one shop variable, and 4M coupon data base with household, shop and time variables. A complex shop database is generated by merging the three databases. The shops are thus symbolic descriptions (see fig. 4).



Fig. 5. Client profile 1

A symbolic hierarchical clustering is carried on these shops. Each class is then characterised by 'shops profiles' through MGS. Each profile is graph-

ically displayed (see fig. 5) through ICM editor. Each profile is a vector providing a 'partial symbolic description' (some variables don't appear in the description) called a 'marking core' in the case of Marking approach.

7 Conclusion

Generalized MGS provides sets of descriptions for a chosen subset, depending on initial discrimination quality request. It can be just a generalization of the whole subset if this quality is null; it may also produce lots of specified descriptions with little extents if the quality is high. MGS could also be used for inference to provide rules if a validation step was added to the supervised learning phase. But its best purpose remains a descriptive process of the data, with generalizing and discriminating potential.

References

- [Alevizos and Morineau, 1992]P. Alevizos and A. Morineau. Tests et valeurs tests. *RSA*, pages 27–43, 1992.
- [Diday, 1988]E. Diday. The symbolic approach in clustering and relative methods of data analysis ; the basic choices. *IFCS*, pages 673–684, 1988.
- [Ganascia, 2000]J.G. Ganascia. Charade et fils : évolution, applications et extensions. *Induction symbolique numérique à partir de données*, pages 303–318, 2000.
- [Gettler Summa *et al.*, 1994]M. Gettler Summa, E. Périnel, and J. Ferraris. Automatic aid to symbolic cluster interpretation. In Springer Ed., editor, *New approaches in Classification and Data Analysis*, pages 405–413, 1994.
- [Gettler Summa *et al.*, 1995]M. Gettler Summa, A. Morineau, and H. Pham ti tong. Marquage des axes et des classes. In *ASU proceedings*, pages 468–472, 1995.
- [Gettler Summa, 1992]M. Gettler Summa. Factorial axis interpretation by symbolic objects. In CEREMADE Ed., editor, *3ème journées numérique- symbolique*, 1992.
- [Gettler Summa, 2000]M. Gettler Summa. Marquages de sous ensembles de données. *Induction symbolique numérique à partir de données*, pages 339–362, 2000.
- [Giordano *et al.*, 2000]G. Giordano, M. Gettler-Summa, and R. Verde. Symbolic interpretation in a clustering strategy on multiattribute preference data. In *Statistica Applicata*, pages 473–495, 2000.
- [Gordon, 1999]A.D. Gordon. *Classification (2nd edition)*. Chapman and Hall/CRC, Boca Raton, 1999.
- [Ho tu *et al.*, 1988]B. Ho tu, E. Diday, and M. Gettler Summa. Generating rules for expert systems from observations. In *Pattern Recognition Letters, Volume 7*, 1988.
- [Massrali *et al.*, 1998]M. Massrali, E. Diday, and M. Gettler Summa. Knowledge extraction from data tables through symbolic data analysis. In Eurostat, editor, *KESDA proceeding*, 1998.

- [Périnel *et al.*, 2003]E. Périnel, A. Ciampi, J. Lebbe, R. Vignes, and E. Diday. Tree growing with imprecise data. In *Pattern Recognition Letters*, pages 787–803, 2003.
- [Stepp, 1984]R. Stepp. A description and user’s guide for cluster/2 a program for conjunctive conceptual clustering. *Report n° UIUCDCS-R61084*, 1984.
- [Vrac and Diday, 2001]M. Vrac and E. Diday. Description symbolique de classes. *Cahier du CEREMADE n° 0115*, 2001.
- [Zighed *et al.*, 1997]D. Zighed, R. Rakotomala, and F. Feschet. Optimal multiple intervals discretization of continuous attributes for supervised learning. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 295–298, 1997.