**Explanatory tools for Machine Learning in the Symbolic Data Analysis Framework**

**[[1]](#footnote-1)**

The aim of this paper is mainly to give explanatory tools for the understanding of standard, complex and big data. First, we recall some basic notions in Data Science: what are complex data? What are classes and classes of complex data? Which kind of internal class variability can be considered? Then, we define “symbolic data” and “symbolic data tables” which express the within variability of classes and we give some advantages of such kind of class description. Often in practice the classes are given. When they are not given, clustering can be used to build them by the Dynamic Clustering method (DCM) from which DCM Regression, DCM canonical analysis, DCM mixture decomposition and the like can be obtained. The description of these class yields by aggregation to a symbolic data table. We say that the description of a class is much more explanatory when it is described by symbolic variables (closer from the natural language of the users), then by its usual analytical multidimensional description. The explanatory and characteristic power of classes can then be measured by criteria based on the symbolic data description of these classes and induce a way for comparing clustering methods by their explanatory power. These criteria are defined in a Symbolic Data Analysis framework for categorical variables, based on three random variables defined on the ground population. Tools are then given for ranking individuals, classes and their symbolic descriptive variables from the more towards the less characteristic. These characteristics are not only explanatory but can also express the concordance or the discordance of a class with the other classes. We suggest several directions of research mainly on parametric aspects of these criteria and on improving the explanatory power of Machine Learning tools. We finally present the conclusion and the wide domain of potential applications in socio demography, medicine, web security, etc.

1. Introduction

A “Data Scientist” is someone who is able to extract new knowledge from Standard, Big and Complex Data. Here we consider complex data as data which cannot be expressed in term of a standard data table where units are described by quantitative and qualitative variables. Complex data happen in case of unstructured data, unpaired samples, multisource data (as mixture of numerical, textual, image, social networks data). The aggregation, fusion and summarization of such data can be done into classes of row units which are considered as new units. Classes can be obtained by unsupervised learning giving a concise and structured view on the data. In supervised learning classes are used in order to provide efficient rules for the allocation of new units to a class. A third way is to consider classes as new units described by “symbolic” variables which values are “symbols” as: intervals, probability distributions, weighted sequences of numbers or categories, functions, and the like, in order to express their within-class variability. For example, “Regions” expressing the variability of their inhabitant, “Companies” expressing the variability of their web intrusion, “Species” expressing the variability of their specimen. One of the advantages of this approach is that unstructured data and unpaired samples at the level of row units, become structured and paired at the classes’ level (see section 2.4).

Three principles guide this paper in conformity with the Data Science framework. First, new tools are needed to transform huge data bases intended for management to data bases usable for Data Science tools. This transformation leads to the construction of new statistical units described by aggregated data in term of symbols as single‐valued data are not suitable because they cannot incorporate the additional information on data structure available in symbolic data. Second, we work on the symbolic data as they are given in data bases and not as we wish that they be given. For example, if the data contains intervals we work on them even if the within interval uniformity is statistically not satisfactory. Moreover, by considering Min Max intervals we can obtain useful knowledge, complementary to the one given without the uniformity assumption. Hence considering that the Min Max or interquartile and the like intervals are false hypothesis has no sense in modern Data Science where the aim is to extract useful knowledge from the data and not only to infer models (even if inferring models like in standard statistics, can for sure give complementary knowledge). Third, by using marginal description of classes by vectors of univariate symbols rather than joint symbolic description by multivariate symbols as 99% of the users would say that a joint distribution describing a class contains often too much low or 0 values and so has a poor explanatory power in comparison with marginal distributions describing the same class. For example, having 10 variables of 5 categories each, the joint multivariate distribution leads to a sparse symbolic data table where the classes are described by a unique bar chart symbolic variable value containing 510 categories and taking for each class 510 low or 0 values. On the other hand, the 10 marginal bar chart symbolic variables value describe the classes by vectors of 10 bar charts of 5 categories each, easy to interpret and to compare between classes. Nevertheless, a compromise can be obtained by considering joints instead of marginal between the more dependent variables.

Symbolic Data Analysis (SDA) is an extension of standard data analysis and data mining to symbolic data. The theory and practice of SDA have been developed in several books ([BOC 2000], [BIL 2006], [DID 2008], [AFO 2018]), many papers (see overviews in [BIL 2003], [DID 2016]), and several international workshops (http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:pa18). Special issue related to SDA has been published, for example in the RNTI journal, edited by, Guan et al., [GUA 2013] on ‘Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis’; in the ADAC journal on SDA, edited by Brito et al. [BRI 2016]; in IEEE Trans Cybern [SU 2016].

This paper is organized in 5 sections. After this introduction the section 2, aims to define symbolic data issued from the descriptions of classes of statistical units (called “individuals”) in order to take care on their internal variability. “Complex data”, “Classes” and “classes of complex data” are defined. The symbolic data appear in the cells of a “Symbolic data table”, where the rows describe classes and the column are associated to variables of symbolic value. Some advantages of symbolic data are finally given in this section.

Section 3 is devoted to the case where the classes are not given, but built by a clustering process. We illustrate this case by two clustering tools: Dynamic Clustering Method (DCM) and by mixture decomposition with the Estimation, Maximization (EM) method. We present different variant of the DCM which can lead to different kinds of clusters depending on the kind of clusters representative: regression, canonical analysis, distributions etc. Then, we show how to build a symbolic data table from the results of these clustering methods. Several criteria measuring the explanatory power of a symbolic data table are suggested. In consequence the explanatory quality of clustering methods can be compared by these criteria.

In section 4 our aim is to define other kinds of explanatory criteria in the case where the initial variables defined on the ground population are of categorical value. We introduce in this case a theoretical framework of SDA based on three random variables. From this framework, we define two kinds of bar chart. The first called “fx(c)” associates to each category x its frequency in the class and the second, called “g c, E (x)” associates its frequency to each event E containing fx(c). These functions yield to characterization of pairs (category, class) by different kinds of criteria. We show that these criteria generalize to symbolic data, the standard Tf-Idf widely used in text mining (see foe example, [ROB 2004]). According to these criteria can be placed in order the individuals, classes, symbolic variables and symbolic data tables from the more to the less characteristic power.

Finally, in the last section 5 we suggest two directions of research. First in this SDA framework different possible parametrization of the criteria expressed in term of concordance or discordance of a class with the other classes are given. An interesting open question is to find in which condition when a sequence of partitions converges toward a trivial partition, such parametric criteria defined on classes converges toward a parametric distribution defined on Ω as it is interesting and economical to obtain from distributions on classes the distribution on the population (in case of concordance or discordance). Second, as explaining for understanding is not discriminating for learning, we suggest a filtering process which improves on a filtered sub population the explanatory power without degrading the discriminating power of any learning machine tool.

2. Introduction to Symbolic Data Analysis

2.1. What are Complex Data?

By definition, “complex data” are any data set which cannot be considered as a “standard statistical units x standard variables” data table. This is the case when data are defined by several data tables with different statistical units and different and unpaired variables coming from multi sources sometimes at multi levels.

Example of complex data in Official Statistics:

The units are REGIONS described by several data tables. For example, each region is described by a first data table where the units are hospitals and the variables are: size of the hospital, number of patients during given periods etc.. In a second data table the units are schools described by the number of pupils, their results in their examinations etc.. In a third data table the units are inhabitants described by socio-demographic variables. In section 2.4 more details are given on that example.

2.2. What are “classes” and “Class of complex data”?

“Classes” are as usual, subsets of any statistical set of units as for example: teams of football players, Region of inhabitant, Level of consumption in health insurance etc.. By definition, a “class of complex data” is a vector of standard classes defined on different statistical spaces of units. For example, in Official Statistics a Region can be considered as a class of complex data denoted CR = (Ch, Cs, Ci) where Ch is the class of hospitals, Cs the class of schools, Ci the class of inhabitants, of this region.

**2.3 Which kind of class variability?**

Classes of statistical units (i.e. individuals”) can express different kinds of variability mainly based on place, time and individuals.Three kinds of variability often happen in practice.First, thevariability between several individuals when time and space are fixed. For example, the variability between the inhabitants of a region at a given period. Second, the variability of a single individual considered at different time and /or place. For example, the performance variability of a team player at different time and/or place. A third one concerns the case where the individual and the time are fixed and the place varies inside the individual. This is the case, for the variability of an individual between its partsas for example,thevariability between the cracks of a cooling tower of a nuclear power plot (see[AFO 2010]).

**2.4. What are “symbolic variables” and “symbolic data tables”?**

The first characteristic of “symbolic variables” is that they are defined on classes. Their second characteristic, is that their values take the variability between the individuals inside these classes into account by “symbols” representing more than only one category or number. Hence, the standard operators of numbers cannot be applied to the values of these kinds of variables, so these values are not numerical: that is why they are called “symbolic” and represented by “symbols” as intervals, bar chart and the like. A “symbolic data table” is a table where classes of individuals are described by at least one symbolic variable. Standard variables can also describe classes by considering the set of classes as a new set of units of higher level.

The Figure 1 is an example of symbolic data table. The statistical units of the ground population are players of French cup teams and classes of players are teams called Paris, Lyon, Marseille and Bordeaux. The variability of the players inside each team is expressed by the following symbolic variables: “Weight” which value is the interval of [min, max] weight of the players of the associated team, “National Country” which value is the list of their nationality, “Age bar chart” is the frequency of the age players being in the intervals: [less than 20], ]20, 25], ]25, 30], ]more than 30], respectively, denoted: (0), (1), (2), (3) in Table 1. The symbolic variable “age” is called “bar chart variable” as the interval of age on which it is defined are the same for all the classes and can therefore be considered as categories. The last variable is numerical as its values for a team is the frequency of the French players in this team among all the French players of all the teams. Hence, this variable produces a vertical bar chart in comparison with the symbolic variable “age” of horizontal bar charts value in Table 1. By adding to the French the same kinds of columns associated with the other nationalities, we can obtain a new symbolic variable whose values are a list of numbers, where each number is the frequency of having players in a team of a nationality among all the players having this nationality among all the teams. A team can also be described by standard numerical or categorical variables as for example, its expenses or the number of goals in a season.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **French Cup teams** | **Weight** | **National Country** | **Age** | **Frequency of French**  **among all French** |
| Paris | [73, 85] | {France, Argentina, Senegal} | {(0) 30%, (1) 70%} | 30% |
| Lyon | [68, 90] | {France, Brazil, Italia} | {(0) 30%, (1) 65%, (2) 5%} | 25% |
| Marseille | [77, 85] | {France, Brazil, Algeria} | {(1) 40%, (2) 52%, (3)8%} | 28% |
| Bordeaux | [80, 90] | {France, Argentina} | {(0) 40%, (1) 60%} | 17% |

**Figure 1.** *An example of symbolic data table where teams of the French Cup are described by three symbolic variables of interval, sequence of categories, “horizontal” bar charts and a numerical variable inducing a “vertical” bar chart.*

This example is built from standard ground data table. In case of complex data we can also built a symbolic data table. For example, National Statistical Institutes (NSI) organize census in their regions on different kinds of populations: hospitals, schools, inhabitants etc. For each region, each of these populations of different sizes are associated to their own descriptive variables. For hospitals: number of beds, doctors, patients, etc.; for schools: number of pupils, teachers, etc.; for inhabitants: gender, age, socio professional category, etc. The regions are the classes of units described by the variable available for all these populations. If we have n regions and N populations (hospitals, schools, etc. for each region), then we get after the symbolic description of each region, a symbolic data table with n rows and p = p1+…+ pN columns where pj is the number of ground variables associated to the jth population. For sure other variables (standard or symbolic) can be added in order to describe other aspects of the regions. Therefore, the unstructured data and unpaired samples at the level of row units, become structured and paired at the classes’ level by a symbolic data table of n rows and p columns.

This example illustrate the importance of complex data in SDA as they constitute a natural and numerous source of symbolic data.

**2.5 Symbolic Data Analysis (SDA)**

The first aim of SDA is to describe classes by vectors of symbolic data in an explanatory way. Its second aim is to extend Data Mining and Statistics to new kinds of complex data symbolic data issued from standard or complex data often coming from the industrial domain. We cannot say that SDA give better results than standard data analysis we can just say that SDA can give good complementary results when we need to work on units which have a higher level of generality and have internal variability. For example, if we wish to know what makes a good player, for sure the data concerns individuals units, but if we wish to know what makes a good team, in this case the natural units are the teams and so, there are classes of individuals.

Moreover, SDA has several advantages. As the number of classes is lower than the number of individuals, SDA facilitates interpretation of results in symbolic decision trees, symbolic factorial analysis etc.. SDA reduces simple or Complex and Big Data. It also reduces missing data and solves confidentiality (as often individuals are confidential but classes are not confidential). It allows adding new variables at the right level of generality.

**3. Symbolic Data Tables from Dynamic Clustering Method and EM**

3.1 **The “Dynamical Clustering Method” (DCM).**

Starting from a given partition P = (c1, …, ck) of a population, this method is based on an alternative use of a representation function g (which associates a representation L to a class c) and an allocation function f which associate a class c to any individual w of the population: C(w) = c in order to improve a given criteria at each step, until convergence.

More precisely,starting from a partition P = (c1, …, ck) of the initial population, the representation function applied to the classes ci produces a vector of representation L = (L1, …, Lk) where g(ci) = Li. A quality criteria can be defined in the following way: where W measures the fit between each class ci and its representation Li . W decreases when this fit increases.

Starting from a partition P(n), the value of the sequence un = W(P(n), L(n)) decreases at each step n of the algorithm. Indeed, during the allocation step an individual w belonging to a class is affected to a new class iff W(P(n+1), L(n)) ≤ W(P(n), L(n) ) = un. Then, starting from the new partition , we can always define a new representation vector where for any i =1 to K, fit best to than or remains unchanged (i.e. . This means: for i =1 to k.

Hence, at this step, we have un+1 = W(P(n+1), L(n+1)) ≤ W(P(n+1), L(n) ) ≤ W(P(n), L(n)) = un. As this inequality is true for any n, this positive sequence decreases and converges.

Moreover, notice that if , then the allocation step consists to change w from one class ci to another class cj when z(w, Lj) < z(w, Li). In this case Lj can be called a “prototype”. Another condition of convergence is that for any c and L, z(c, g(c)) ≤ z(c, L), in that case g(c) is an “optimal prototype”.

**3.2 Examples of DCM applications**

The classical k-means method is the case where the Lk are the means of the Ck. When in the DCM the Lk are probability densities, we have a mixture decomposition method (see, [DID 75], [DID 2005]) which improves the fit (in term of likelihood) between each class (of the partition) and its associated density function. More precisely, in this case each individual is associated by the allocation function to the density function of highest value for this individual. In case of representation by a regression, each individual is allocated to the class C’i  if this individual fit the best the regression Li, (see [CHA 77]), more generally in case of representation by canonical axis see [DID 78], [DID 86]. There are many other possibilities such as when representation of any class can be a distance [DID 77], a functional curve [DID 76], points of the population [DID 73], a factorial axis [DID 72].

DCM Canonical analysis is a general method which aim (see Figure 2) by giving p blocs of variables, is to find simultaneously k classes of individuals and m canonical axis fitting the best. The mathematical question as expressed in [DID 86] is settled as follows:

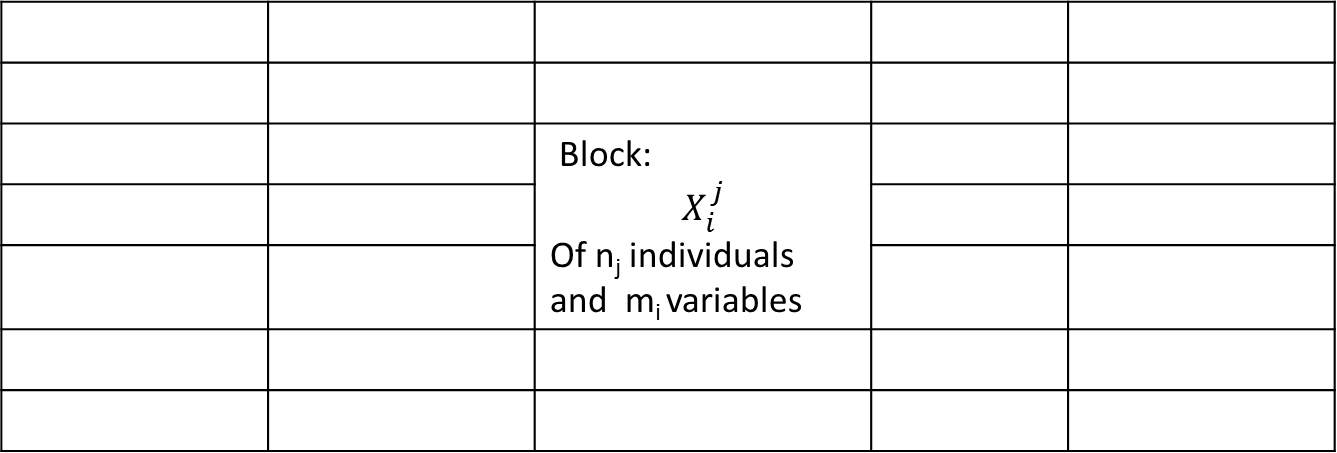
Maximize: W(P, ξ, Z) = , >

Where the are linear combinations of the variables of the ith block of variables for the jth class of individuals associated to the jth canonical and the 𝑙th axes among the m axis.

n individuals

n individuals

P1 P2 Classe Pj Pk



p blocks

of variables

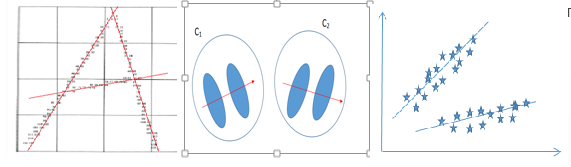
**Figure 2.** DCM Canonical analysis of k blocks of individuals described by p blocks of variables

Notice that the DCM Canonical Analysis contains as a special case:

* DCM Principal Component analysis [DID 72], (see figure 3 a).
* DCM regression [CHA 77], (see figure 3 c)

In case of categorical variables it leads in [DID 78] to:

* DCM Factorial correspondance analysis
* DCM discriminant analysis (see figure b).



**(a) (b) (c)**

**Figure 3.** DCM PCA: find simultaneously classes and first axes of local PCA which fit the best

DCM Discriminant Analysis: find simultaneously classes and first axes of local factorial discriminant analysis which fit the best.

DCM Regression: find simultaneously classes and local Regressions which fit the best.

. For an overview on DCM see [DID 79], [DID 80].

**3.3. Clustering methods by mixture decomposition**

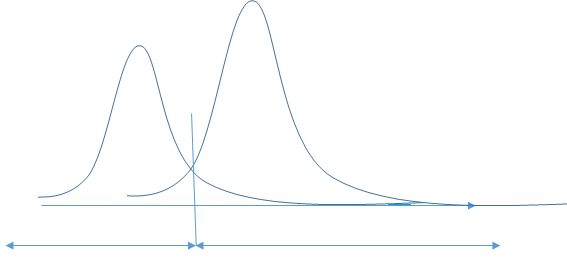
In case of mixture decomposition by DCM ([DID 75] , [DID 2005] ) for partitioning or EM [DEM 77]), for fuzzy partitioning, the joint probability densities are associated to each obtained cluster. More precisely**,** the DCM aims tobuild iteratively a partition Pi  and simultaneously a probability density Li in a dynamical clustering process. The obtained partition maximizes iteratively the following criteria denoted W where w denote the likelihood of the Li for the Pi such that:

**.**

The EM method aims to obtain a fuzzy partition maximizing the likelihood of the probability density f where the are the probability densities of the mixture decomposition satisfying at the individual w with the parameters ai a given model, such that:

f(w, a) = with

Notice that the EM obtained fuzzy partition fit the obtained probability densities, but the exact partition is defined by and do not fit the associated probability density. As shown in Figure 4 in case of a two class’s partition, the classes does not fit their associated probability densities. At opposite, the exact partition given by DCM fit its obtained probability densities but do not fit its associated fuzzy partition. Therefore, both methods can gain to be used alternatively in order to improve their obtained partition (exact or fuzzy).



Class P 1

Class P 2

Figure 4. *The Exact Partition induced from the density functions obtained from EM does not fit these obtained density functions. This is not the case at the convergence of DCM where the obtained partition and its associated density functions fit exactly..*

**3.4 Symbolic Data Tables from clustering**

Building a symbolic data table where each unit are the obtained clusters can be done in three ways: directly if the obtained clusters define a partition, from the marginal induced by the joint distribution associated to each cluster provided by EM or DCM, or from the membership weight of the individuals if we have fuzzy clusters as in EM mixture decomposition.

If Lm is the representative of the class cm, then the weight tk(wi) of an individual wi in class ck is given by: where d is the dissimilarity used by the clustering method which has produced the classes. Then, the histogram for the mth class and the jth variable is given by:

(1)

Where is the value taken by the variable Xj for the individual wi. In other words Xj(wi.) = .

, …, ) is a vector of Dirac mass defined on V intervals partitioning the domain Dj of the numerical variable Xj such that: takes the value 1 if and 0 elsewhere. When the are categorical values instead of intervals, we obtain a bar chart and takes the value 1 if is the category and the value 0 elsewhere.

When instead of a fuzzy partition as the one given by EM we have an exact partition denoted as the one induced by or directly by DC, we can build in the same way an histogram or a bar chart by setting:  
, for any and .

In SDA, in order to increase the explanatory power of the obtained symbolic data table, first the chosen number of intervals is preferably chosen not numerous (about 5, but it can be increased if needed), second the size and position of these intervals can be obtained in an optimal way in order to maximize the distance between the symbolic description of the classes (see [DID 2013]a), third weights can be added to the variables when the clustering method use DCM with cluster representative as PCA, regression, canonical analysis, multi blocks approach as in [BOU 2017, 2018], by considering that the categories of each symbolic bar chart variables constitute the blocs of standard numerical variables.

By this way from any clustering method we can obtain a symbolic data table on which SDA can be applied.

The SDA aim is to study the obtained symbolic data table in order to get complementary knowledge and more explanatory results than the usual standard interpretation. At the contrary, in standard mixture decomposition the description of each class is often just given by the analytical expression of the joint probability density fi associated to each class. For example, in case of Gaussian model, the joint is described by a big correlation matrix very heavy to interpret when the number of variables are numerous. A more explanatory way is to describe the joints fi associated to each class Ci by their marginal fij. These marginal associated to each class can then be described by several kinds of symbolic data as histograms or interquartile intervals or any kind of symbolic data. The three steps of describing clusters from the less towards the more explanatory are summarized in the Figure 5. Notice that the obtained symbolic data table is not only by itself a complementary way to interpret the results of the mixture decomposition but moreover it is a starting point for applying the tools of SDA as a symbolic PCA or many other kinds SDA tools on the obtained clusters or fuzzy clusters and their symbolic description in order to enhance considerably their interpretation.

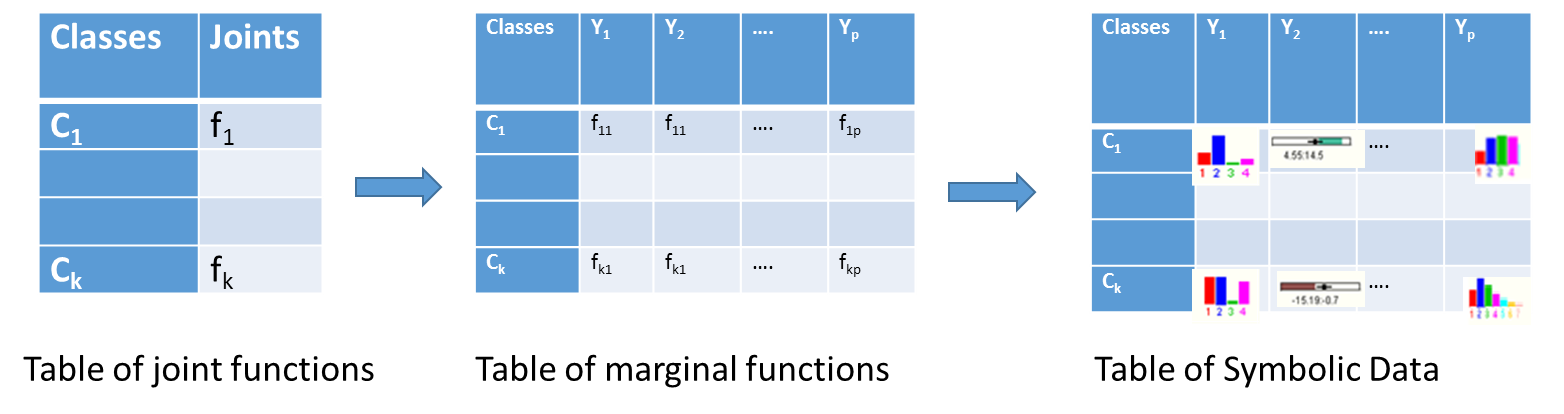


Figure 5: *After a mixture decomposition from the less to the more explanatory way of the obtained clusters.*

**3.5. A general way to compare results of clustering methods by the “explanatory power” of their associated Symbolic Data Table.**

There are several way for building a symbolic data table from a given set of classes. The best way in SDA is to get a meaningful symbolic data table by maximizing the discrimination power of the symbolic data associated to each variable for each class. A discrimination degree can be calculated by a normalized sum (to be maximized) of the dissimilarity two by two between the symbolic descriptions. Such kind of dissimilarities can be found in [DEC 96], [DEC 2006], [BOC 2000], [DEC 2006], [DID 2008], [ICH 94], [GOW 91], [IRP 2008]. In case of histogram value variables an example of discriminating tool is given in [DID 2013]a . In summary, there are at least two (related) ways to obtain a meaningful symbolic data table:

. distances between rows to be maximized,

. entropy in each cell of the symbolic data table to be minimized. More details are given in [DID 2019].

**3.6 Quality Criteria of Classes and Variables based on the cells of the Symbolic Data Table containing intervals or inferred distributions**

Since long time ago, much work has been done on robust intervals which can be useful to measure the robustness quality of a symbolic interval-valued variable, see [ROY 86], [HOR 98]. The quality of the inferred distributions contained in the cells of the obtained symbolic data table can be also measured by classical model selection criteria like Bayesian Information Criteria (BIC), Minimum Description Length (MDL), Akaike’s Information Criteria (AIC), Minimum Message Length (MML), or other criteria of this kind based on the likelihood estimation.

In the coming section 4 other kinds of explanatory power of clustering methods based on the symbolic data table that they induce by aggregation are given.

**4. Criteria for ranking individuals, classes and their bar chart descriptive symbolic variables**

**4.1. A theoretical framework for SDA**

A general theoretical framework for SDA can be found in[EMI 2018]. Here we define a theoretical framework for SDA in case of bar chart symbolic variables. Let be three random variables C, X, A defined on the ground population Ω in the following way:

C a class variable: Ω 🡪 P such that C(w) = c where c is a class of a given partition P.

X can be considered as a vector of variables and X(w) a vector of categories (called “metabins” see [Did 2013]b) containing one category for each variable. In this case X is a mapping from Ω in M the set of metabins. For sure it can be interesting to study the joint explanatory power of such metabins for a given class. Nevertheless, following the principle that the explanatory power of a metabins is the sum of the explanatory power of each of its bin, we use in the following part of this paper the marginal case by considering X as a unique variables. In the following all the criteria express marginal explanatory power that is why CX is defined in the following way:

X a variable: Ω 🡪 M such that X(w) = x is a category among a set of categories M.

A is an aggregation function which associate to a class c a symbol s = A(c) which can be a min, max interval or an interquartile interval or a cumulative distribution or a quantile function or a bar chart, and the like issued from an aggregative process on individuals. Here, s is restricted to be a mapping from M to [0, 1].

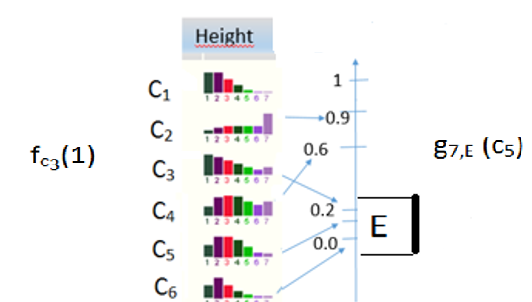
From C, X and A, we can build the so called "symbolic random variable" SH which is defined on Ω and takes its values in [0, 1],such that SH associates to w Є Ω the value SH(w) = A(C(w))(X(w)).

In the following we consider that M is restricted to be a set of categories and f c: M 🡪 [0, 1] is the bar chart induced by the class c = C(w). Then, if x = X(w), we have SH(w) = f c(x) where f c = A(c) is a bar chart symbol. This symbol f c is “Horizontal “ in Fig 6, that is why H is the index of S.

We can also define another bar chart g x: C 🡪 [0, 1] such that if c = C(w) and x = X(w) then

gx(c) = {w’ Є Ω / f C(w’)(x) = f c(x) }| / |Ω|. If v (x) = {f c(x)/ c Є P}. We can now define another symbolic random variable Sv: SV(w) = g x(c) where A(v(x)) = gx is a bar chart symbol.

More generally, we can define g x, E (c) = |{w’ Є Ω / f C(w’)(x) Є E}| / |Ω| = SVE(w) where E is an interval included in [0, 1] which generalizes the preceding case where E was reduced to f c(x). These functions are illustrated by an example given in Fig 6.



**Figure 6** A symbolic data table reduced to 6 classes among many others and to a unique symbolic variable X = Height of bar chart value (with 7 categories), for each class. Each class ci is associated to a bar chart fci and we represent the gxE(c) value for the category x = 7, the interval event E around fc5(7) = 0.2 and for the class c5

**4.2 Characterization of a category and a class by a measure of discordance.**

We say basically that a category x is “characteristic” of a class c if it is frequent in the class and rare in the other classes. In other words, there is a “discordance” between the class c and the other classes for this this category x. More generally, if we denote E an interval contained in [0, 1] such that fc(x) belongs to E, the category x is “characteristic” of the class c for this event E if fc(x) is large and rarely fc’(x) belongs to E when c’≠ c varies among all the classes of the partition P.

A characterization criteria W of a category x and a class c can be measured by:

W(x, c) = f c(x) / gx,E (c) or by W(x, c) = - f c(x) Log( gx,E (c)) (many other variants are possible like in the case of the Tf-idf criteria even if this criteria is very different from the standard Tf-idf)

In order to have a criteria varying between 0 and 1 we can use:

**W(x, c) =**  f c(x) **/ (1+** gx, E (c)**),**

Moreover, given an event E, both criteria W express, how much a category x is “discordant” of a class c versus the other classes c’ of the given partition P. This criteria means that a category x is even more discordant of a given class c and for an event E, its frequency in the class c is large and the proportion of individuals w taking the x category in any class c’ and such that fc’(x) belongs to the event E(x, c) is low in the ground population Ω.

Giving x and c, several choices of E can be interesting. We give now four examples of events E.

For a characterization of x and c in the neighborhood of f c(x):

E1= [f c(x) – ε, f c(x) + ε] for ε > 0 and f c (x) Є [ε, 1- ε].

For a characterization of the higher values than f c(x): E2 = [f c(x), 1].

For a characterization of the lower values than f c(x): E3 = [0, f c(x)].

In order to characterize the existence of the category x: E4 = ]0, 1].

Hence, a category x is characteristic of a class c when it is frequent in the class c and rare to appear :

in the classes c’≠c with a frequency in a neighborhood of fc(x) if E = E1, with a frequency above (resp. under f c(x)) if E = E2  (resp. E = E3), with a frequency strictly higher then 0 (i.e. rare to appear in classes c’ different of c) if E = E4. In all these cases gx, E (c) is low and so W is high if f c(x) is high.

**Singular, typical or specific characteristic discordance of a category**

When w such that X(w) = x and C(w) = c varies in Ω, there are four cases to consider depending on the fact that the category x is frequent (denoted Fc) or rare (denoted Rc) in c and this category x is frequent (denoted FE) or rare (denoted RE) in the set of classes c’≠ c such that f c’(x) belongs in E, Hence, we have four cases called Fc FE,

Fc RE, Rc FE , Rc RE depending on x or c’, the cases Fc FE and Rc RE cannot give any specific value to W(x, c). The case Fc RE where a category is frequent in c (i.e. f c(x) high) and its frequency inside the classes c’≠c rarely appear in E (i.e. g x(c) low), leads to a value of W(x, c)close to 1. In this case, we can say that the category is “typical” of c and the criteria W measures its typicality. In the case Rc FE where the category is rare in c (i.e. f c(x) low) and has a frequency in the classes c’≠c frequently inside E (i.e. g x(c) high), leads to a value of W(x, c) close to 0. In this case, we can say that the category is “singular” in comparison with the other classes and the criteria W measure the singularity of the category.

The categories x which satisfies a value W(x, c)close to 0 or 1 can be said “specific” to c as it is typical or singular. Therefore, we can say that a category x of a variable Y is specific of c if W(x, c)LogW(x, c) is low. We can also say that in comparison with the other classes the class c is “discordant” for the category x. We can also say that the explanatory power of a couple (x, c) is 1- W(x, c)LogW(x, c).

**4.3 Link between a characterization by the criteria W and the standard Tf-Idf**.

The basic idea of the Tf-Idf is to characterize a category of a class by the fact that it is frequent inside the class and rare in the other classes of the given partition P.

In other words, if n(x, c) is the number of occurrences of x in c, K is the number of classes and k(x) the number of classes containing x, then the Tf-Idf of x and c can be written:

Tf-Idf (x, c) = (n(x, c)/ |c |) (K/k(x))

Hence, the Tf-Idf is even greater than x appears in a class and rarely in the other classes.

**Proposition**

If the classes c of the partition P have the same size and their elements w are either all such that: f c(x) = 0, neither all such that : f c(x) > 0 then W(x, c, ]0, 1]) is the standard Tf-Idf for the value (x, c).

**Proof**:

By definition:

n(x, c)/ | c | = f c(x) (1)

So from the hypotheses that all the classes has same size results we get: K |c| = |Ω|. From the hypotheses that the elements of the classes take all the value x or all takes another value, we get: k(x) |c | = {w/ f C(w)(x) > 0}. Therefore

K / k(x) = (K |c|) / (k(x) |c|) = |Ω| / {w/ f C(w)(x) > 0}. So finally:

K / k(x) = 1/ g x,]0, 1] (c) (2)

Therefore: Tf-Idf (x, c) = (n(x, c)/ | c |) (K/k(x)) implies from (1) and (2):

Tf-Idf (x, c) = f c(x)/ g x,]0, 1] (c) = **W(x, c)**

End of proof.

Notice that there are several other closed way to define the Tf-Idf. For example, by using a Log as followed:

Tf-idfLog (x, c) = (n(x,c)/ | c |) Log((K/k(x)))

In this case by setting: WLog **(x, c) = -** f c(x)Log(g x,]0, 1] (c)), we obtain:

Tf-IdfLog(x, c) = f c(x) Log(1/ gx, ] 0, 1] (c)) = **W**Log**(x, c,** **]0, 1]))**

More generally, if Tf-idf ’ (x, c) = Tf(n(x,c)/ | c |) Idf(K/k(x)), where Tf and Idf are other ways to define the Tf-idf. Then, by setting W’**(x, c) = Tf(**f c(x))**/** Idf(g x,]0, 1](c)) and by using (1) and (2) we obtain: Tf-Idf’(x, c) = **W**’ **(x, c,** **]0, 1]).**

**Another kind of SDA characterization criteria closer from the tf-idf :**

It is based on another choice of g by g’:

g’ x, E (c) = |{c Є P / f C(w)(x) Є E}| / |Ω|

By setting: W’(x, c) = f c(x) / g’x,E (c) , W’(x, c) = - f c(x) Log( g’x,E (c)) or W’(x, c) = f c(x) / (1+ g’x, E (c)), We get a characterization criteria equal to variants of the tf-idf when E = ]0, 1] (i.e. E = E4).

With this criteria W’, the typicality, singularity or specificity can be calculated in the same way then with W but with a different meaning as in this case the category x appears in a frequent (denoted FE) or rare (denoted RE) number of classes c’≠ c such that f c’(x) belongs in E.

**Other criteria of characterization**

Other kinds of characterization criteria can be used. The popular ‘test value,’ developed in [LEB 84], may also be used to measure a characterization of a category in a bar chart contained in a cell. The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. A simple way can be the ratio between the frequency of a category in a class and the mean of the frequencies of the same category in all the classes of the given partition.

**4.4 Ranking the individuals, the symbolic variables and the classes of a bar chart symbolic data table**

A bar chart symbolic data table is defined by a set of p symbolic variables Xj in column describing in rows K classes ck and containing in each cell a bar chart of categories denoted zjm for m = 1 to mj associated to each variable Xj. In the following, the characterization criteria W covers all kinds of variants including W’. Each category associated to each cell of a bar chart symbolic data table can be characterized by a W criteria. Giving an individual w, we have its class C(w) = ck and its categorical value Xj(w) = zjm for any variable Xj . Then, by summing on the characterization W value of the category zjm for j = 1 to m row associated to the class ck, we get a characterization of this individual w. Then, by summing on the characterization W value of a selected category in all the cells of each row (resp. column) of this symbolic data table, we obtain a characterization of each class (resp. variable). In the same way, in summing on the characterization W value of all the cells, we can obtain a characterization of the symbolic data table. By that way, we can find a typical, singular or specific ranking of individuals, classes, bar chart variables or symbolic data table.

For example, the characterization measure of an individual w for the jth variable such that Xj (w) = zjm, C(w) = ck for the event E, is defined by:

**W(**zjm , ck**) =** f c(zjm)**/** (1+g x, E (ck)).

Therefore, the characterization measure of an individual w can be:

CI(w) = ∑j = 1, p W(Xj(w), C(w)).

We can then define a typicality measure of a symbolic variable Xj by:

CV(Xj) = ∑k = 1, K Max m = 1, mj W(zjm , ck).

We can also define a typicality measure of a class c by:

CC(c) =∑j = 1, p Max m = 1, mj W(zjm, c).

We can finally define a typicality measure of a partition P (which can be called “typicality of the symbolic data table” defined by the symbolic class descriptions), by:

CT(c) =∑k = 1, K CC(ck)

The singularity measure can be calculated by using the min instead of the max.

**Ranking:**

We can then place in order from the less to the more characteristic the individuals w, the symbolic variables Xj for j= 1, to p and the classes ck for k = 1 to K, by using respectively the CI, CV , CC and CT characteristic measures. These orders are respectively denoted OCI, OCV, OCC and OCT. Notice that in all these cases, we can associate a metabins to each individual or class by choosing the most characteristic category of each variable. A hierarchical or pyramidal clustering on these metabins can facilitate the interpretation of the most characteristic individuals or classes as they can be close in the ranking but for different reasons (expressed by different metabins). For example, in a study on the Cause-Specific Mortality in European Countries [AFO 2018], Bulgaria, Romania with Singularity Level (SL) respectively equal to 271 and 252, was the most discordant until Poland (SL = 152) with high level of mortality mainly for circulatory problem. Denmark (SL= 152) and France (SL = 149) was close from Poland in the ranking but for very different reasons with low level of mortality.

From these basic criteria (CI, CV, CC and CT), many other can be considered, for example by using the mean or the sum or the median values instead of the Max or the Min. Also, instead of giving the same weight to the categories in the sums, we can give a different weight as the one obtained by a DCM canonical analysis given in [DID 86] or more recently in [BOU 2017, 2018], by considering that the categories of each symbolic bar chart variables constitute the blocs of standard numerical variables.

**5. Two directions of research**

**5.1 Parametrization of concordance and discordance criteria**

Notice first that instead of considering the characterization criteria W which express a “discordance” between a class and the other classes, we can consider a criteria which express a “concordance”. This criteria denoted Wconc has many variants nevertheless it can be written basically (with E given fixed) in the following form:

Sconc(w) = f C(w)(X(w))gx,E (C(w)) . Therefore:   
Sconc (w) = SH(w) SV(w)

This criteria varies between 0 and 1 and express a concordance between the class C(w) and the other classes as it highest value is obtained for individuals w such that the frequency of the category X(w) in C(w) is high and simultaneously the number of individuals in the other classes (for W) or the number of other classes (for W’) having a high frequency for X(w) is also high.

The discordance can be written in the same way:

Sdisc(w) = f C(w)(X(w))/ (1+gx,E (C(w)) . Therefore:   
Sdisc (w) = SH(w)/(1+ SVE(w))

**Proposition**

When the partition P is the trivial partition (i.e. C(w) = {w}) and X(w) = x, then we have: Sconc(w) = fE(x) and Sdisc(w) = 1/(1+ fE(x)) where fE is the frequency of individuals w such that f(X(w)) Є E.

**Proof**

In the case where C(w) = {w} and X(w) = x, we have: f C(w)(X(w)) = f {w}(x) = 1 as it is the frequency of the category x in a class containing only w which associated category is x. Moreover, we have :

gx(c) = {w’ Є Ω / f C(w’)(x) = f c(x) }| / |Ω|

= {w’ Є Ω / f {w’}(x) = f c(x) }| / |Ω|

= {w’ Є Ω / f {w’}(x) = 1}| / |Ω|.

= f(x) the frequency of x in the ground population Ω

Therefore, gx(c) becomes the frequency of x in Ω. In the same way gx, E(c) becomes the frequency denoted fE(x) of E. Hence, we obtain when P is the trivial partition (i.e. C(w) = {w}) and X(w) = x:

Sconc(w) = f C(w)(X(w))gx,E (C(w)) = fE(x).

Sdisc(w) = f C(w)(X(w))/ (1+gx,E (C(w)) = 1/ (1+ fE(x))

These results leads to an explanatory interpretation of Sconc and Sdisc  in the case of such trivial partition, as we can say that the concordance of an individual will be all the greater in that the frequency of its category inside the interval E is great too. In the same way, we can say that the discordance of an individual will be all the greater in that the frequency of its category in E is small.

The parametric case:

If SH(resp (SV) depends on a parameter “a “ (resp “b”) under some models assumption (Multinomial , Dirichlet or the like), and having a sample of Ω: {w1,…, wn} , we have:

Sconc (w, a, b) = SH(w , a) SV(w , b)

Sdisc (w, a, b) = SH(w, a)/(1+ SV(w, b))

The parameters, a, b can be estimated by maximizing the following likelihood

Lconc (Sconc; a, b) = ∏i=1,n Sconc (wi , a, b)

In the same way we can parametrize a parametric discordance by maximizing:

Ldisc(Sdisc; a, b) = ∏ i=1,n SH(wi, a) / (1+ SV(wi, b))

If we define a law F on the fc and a law G on the g x,E(c), a more accurate parametrization can be settled in the following way : find a’, b’, a, b which maximizes:

Lconc (Rconc; a’, b’, a, b, E) = ∏ i=1,n RE (wi , a’, b’)SH(wi , a) SV(wi , b) where:

Rconc (wi , a’, b’) = F(fC(wi); a, a’)G(gX(wi),E ; b, b’).

Ldisc (Rdisc; a’, b’, a, b, E) = ∏ i=1,n RE (wi , a’, b’)SH(wi , a) / (1+ SV(wi, b)) where:

Rdisc(wi , a’, b’) = F(fC(wi); a, a’)/ (1+G(gX(wi),E ); b, b’)).

Like in 4.4 the orders OCI, OCV, OCC and OCT can be used with all these kinds of criteria**.** An interesting open question is to find in which condition when a sequence of partitions converges toward the trivial partition, the parametric concordance Sconc(w, a, b) converges toward a parametric frequency fE(X(w), a,b)) on Ω of same parameters. Same question for the discordance case. Another open question is to extend the concordance and discordance to the case where X is a numerical random variable and the symbolic value variables are distributions. Much has also to be done by considering joint explanatory criteria instead of marginal as it has been done in this paper as in case of dependencies the results of both approaches can be very different. A compromise can be to apply DCM canonical analysis and to use the features induced by the best explanatory canonical axis.

**5.2 Improving the explanatory power of any machine learning tool by a filtering process**

Explaining for understanding is not discriminating for learning (see [DID 2018]). Our aim is now to give a filtering process which improves on a filtered sub population the explanatory power without degrading the discriminating power of any learning machine tool. We suppose here that we have already obtained a clustering from a basic sample where the predictive values are known in case of supervised data.We have to consider two cases depending on the fact that the data are supervised or not.

In case of unsupervised data we have to allocate new individuals to the best fitting representative associated to each cluster. For example, in the case of the k-means, we associate any new individual to the cluster of closest mean. In the DCM case where the class representative is a distribution (like in DCM Mixture decomposition see [DID 75], [DID 2005]), any new individual is allocated to the cluster which associate density function, maximizes the likelihood of this individual. For any individual and in any case we can obtain an order of preference of the clusters from the best fitting representative to this individual to the less representative. Hence, by this way, from any individual we can place the clusters in an order denoted O1.

In case of supervised data, there are two steps. In the first step the aim is to allocate a new individual (which predictive value is not given) to the best cluster. In the second step the aim is to obtain the predictive value of this new individual, from the local model associated to this cluster. For example, if we allocate a new individual to a cluster modeled by a local regression given by a DCM regression (as in [CHA 77]), we can then obtain its predictive value by using this regression. The same can be done if instead of having a local regression, we have a local decision tree, a local SVM, a local neural network, etc... In order to find the best new individual cluster allocation, we can only use the given data without the predicted value variable as for the new individuals for sure this value is not given. Coming back to the basic sample where now the predicted value variable associate to each individual is its cluster, we can use, a supervised machine learning tool (SVM, Neural Network or any black box learning machine method) on these data. In that way, any new individual can be associated to a preference order of the clusters from the best allocation to the worse. Hence, by this way, an individual can place the clusters in an order denoted O2.

We can also associate to any new individual its fit to the symbolic description associated to any obtained cluster. For example, in the numeral case, if the symbolic descriptions are density functions fj, we can use the likelihood product of the fj(xj) for j = 1, p where xj is the value taken by this individual for the jth initial variable. We can then place in order the clusters from their best to the lower fit to this individual. We can also replace fj(xj) by W(xj, c) in the categorical case. Hence, by this way, an individual can place the clusters in an order denoted O3. O3 is an explanatory order as it is based on the symbolic description of the clusters.

Finally given a new individual, we can place in order the obtained clusters by three ways: O1, O2, O3. Several strategy (see here under) are then possible. Having chosen one of them we can continue the machine learning process: we allocate the new individual to a cluster and then adding it to this cluster, then finding a best fit representative and so on until the convergence of DCM until a new partition and its associated local models.

Machine learning filtering strategies:

The idea is to add (i.e. to filter) a new individual to the cluster and to the aggregation process leading to a new symbolic description, if it improves simultaneously at best the fit between the cluster and its representative and the explanatory power of its associated symbolic description.

The first kind of filtering strategy is to continue the learning process with only the individuals which have at best position the same first cluster (i.e. same leader) in the order O1, O2 and O3. Another kind of strategy is to continue the learning machine process with only the individuals whose clusters at best position (in the three orders) are not beyond then a given rank k. Then the individual is allocated to the cluster of best rank following O1 , O2 or O3 alternatively or depending if you wish more explanatory power or better decision. Other strategies are also possible by adding OCI, and (or) OCL to the orders O1, O2 and O3 or any given order on the individuals a priori given by an expert. It is also possible to reduce the number of variables by selecting the first ones in the OCV order or by any learning machine method aiming to select variables. In any filtering strategy, the learning process progress with individuals which improve the explanatory power of the machine learning as much as possible without degrading at all or not much the efficiency of the obtained rules. When a sub-population is obtained, the process can continue with the remaining population and leads to other subpopulations and so on when the population increases or until the whole population has been considered.

**Conclusion**

The aim of this paper was to give tools related to the part of our brain needing to understand what happen and not to the other parts of our brain needing to take efficient and quick decision without knowing how (for example for face recognition). Classes obtained by clustering or a priori given in unsupervised or supervised learning machine are here considered as new units to be described in their main facets and to be studied by taking care on their internal variability. We have shown that classes can be obtained by DCM which have many variants depending on the kind of representative (means like in the popular k-means clustering but more generally principal components, regressions, distributions, canonical axis, etc.). Then, we have given tools for building symbolic data describing these classes on which SDA methodology can be applied. We have focused on bar chart symbolic descriptions of classes but for sure other kind of symbolic representation of classes can be done in the spirit. Several explanatory criteria has been defined from which individuals, classes, symbolic variables and symbolic data tables can be placed in order from the more towards the less characteristic. Much remains in order to compare and improve the different criteria and to extend them into the parametric and numerical cases. These tools have potential applications in many domains. For example, in order to compare the explanatory power of clustering algorithm or more generally for improving the explanatory power of machine learning. In practice these criteria can be applied in order to find discordant countries of Europe on political questions or on illnesses and death causes, to place in order power point cooling towers, to find boats of risk in a harbor, to find singular stocks behavior, to find typical text section of a book or specific web intrusions, to find discordant or concordant images, etc. These explanatory tools have an immense potential for research and applications.

**References**

[AFO 2018] AFONSO F., DIDAY E., TOQUE C. “Data Science par Analyse des Données Symboliques”. *Book (448 pages). TECHNIP editor*.

[AFO 2018] AFONSO F., DOLINAR A.L., KORENJAK-CERNE S., DIDAY E. “Analysis of Gender-Age-Cause-Specific Mortality in European Countries with SYR Software”. Brito, Dias (Editors) Symbolic Data Analysis Workshop SDA 2018. https://sda2018.wixsite.com/ipvc

[AFO 2010] AFONSO F., DIDAY E., BADEZ N., GENEST Y., “Symbolic Data Analysis of Complex Data: Application to nuclear power plant”. *COMPSTAT’2010 , Paris*

[BIL 2003] BILLARD L., DIDAY E., "From the statistics of data to the statistic of knowledge: Symbolic Data Analysis". JASA. *Journal of the American Statistical Association. Juin,* *(2003),* *Vol. 98, N° 462.*

[BIL 2006] BILLARD L., DIDAY E., “Symbolic Data Analysis: Conceptual Statistics and Data Mining. *Wiley Series in Computational Statistics. Chichester: Wiley; 2006, 321. ISBN: 0-470-09016-2.*

[BOC 2000] BOCK HH, DIDAY E., “Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data”. *Heidelberg: Springer-Verlag; 2000, 425. ISBN: 3-540-66619-2.*

[BOU 2018] BOUGEARD, S., CARIOU, V., SAPORTA, G. & NIANG, N., “Prediction for regularized clusterwise multiblock regression”. *Applied Stochastic Models in Business and Industry, 1-16*.

[BOU 2017] BOUGEARD, S., ABDI, H., SAPORTA, G. & NIANG, N., “Clusterwise analysis for multiblock component methods”. *Advances in Data Analysis and Classification. 12(2), 285-313.*

[BRI 2016] BRITO P, NOIRHOMME-FRAITURE M, ARROYO J., “Special issue on symbolic data analysis”. *Editorial. Adv Data Anal Classif 2015, 9:1–4.*

[CHA 77] CHARLES C. "Régression typologique et reconnaissance des formes", *Thèse de 3ème cycle, Juin 1977, Université Paris IX-Dauphine and INRIA Rocquencourt 78150 (France).*

[DEC 96] DE CARVALHO FAT. “Extension based proximity coefficients between constrained Boolean symbolic objects”. *In: Hayashi C et al., eds. Proceedings of IFCS’96. Berlin: Springer-Verlag; 1998, 370–378.*

[DEC 2006] DE CARVALHO F, SOUZA R, CHAVENT M, LECHEVALLIER Y. “Adaptive Hausdorff distances and dynamic clustering of symbolic interval data”. *Pattern Recogn Lett 2006, 27:167–179.*

[DID 75] DIDAY E, SCHROEDER A. "A new approach in mixed distributions detection". *E. Diday, A. Schroeder. RAIRO. Vol. 10, n°6.*

[DID 78] DIDAY E. "Analyse canonique du point de vu de la classification automatique". *Rapport Laboria (Avril 1978) n°293. INRIA, 78150 Rocquencourt, France.*

[DID 86] DIDAY E. "Canonical analysis from the automatic classification point of view", *Control and Cybernetics, 1986 Vol. 15, n°2.*

[DID 2019] DIDAY E. “Pouvoir explicatif et discriminant de variables et de tableaux de données symboliques”. *To appear in RNTI journal. Contact:* [*info@editions-rnti.fr*](mailto:info@editions-rnti.fr)*.*

[DID 2016] DIDAY E. “Thinking by classes in data science: symbolic data analysis”. WIREs *Computational Statistics Symbolic Data Analysis Volume 8, September/October 2016. Wiley Periodicals, Inc. 191.*

[DID 2013]a DIDAY E, AFONSO F, HADDAD R. “The symbolic data analysis paradigm, discriminant discretization and financial application”. *In: HDSDA 2013 Conference, Beijing, China.In RNTI-E-25. Paris: Hermann; 2013, 1–14.*

[DID 2013]b DIDAY E. “Principal component analysis for bar charts and Metabins tables.” *Stat Anal Data Mining 2013, 6:403–430. doi:10.1002/sam.11188.*

[DID 73] DIDAY E “The Dynamic clusters method in non-hierarchical clustering” *International journal of Computer and information Science Vol 2, n° 1, 1973. DOI: 10.1007/BF00987153.*

[DID 72] DIDAY E. "Introduction à l'Analyse factorielle typologique". *Rapport Laboria n°27, (1972). INRIA, 78150 Rocquencourt, France.*

[DID 2005] DIDAY E., VRAC M. **"**Mixture decomposition of distributions by Copulas in the symbolic data analysis framework" . *Discrete Applied Mathematics (DAM). Volume 147, Issue 1, 1 April (2005), Pages 27-41.*

[DEM] DEMPSTER A, LAIRD N, RUBIN D. “Maximum likelihood from incomplete data with the EM algorithm.” *J R Stat Soc Series B Stat Methodology 1977, 39:1–38.*

[EMI 2018] EMILION R., DIDAY E. " Symbolic Data Analysis Basic theory " *Chapter in Advances in Data Sciences, edited by Prof. Saporta, Prof. Wang, Prof. Diday, Prof Rong Guan,2018, Publisher: ISTE-Wiley.*

[GUA 2013] GUAN R, LECHEVALLIER Y, SAPORTA G, WANG H. “Advances in Theory and applications of High Dimensional and Symbolic Data Analysis*”,* *vol. E25. Hermann, MO: RNTI (2013).*

[GOW 91] GOWDA KC, DIDAY E. “Symbolic clustering using a new dissimilarity measure”. *Pattern Recogn 1991, 24:567–578. 24:567–578.*

[HOR 98] HORN S, PESCE AJ, COPELAND BE. “A robust approach to reference interval estimation and evaluation. *Clin Chem”. 1998, 44:622–631.*

[ICH 94] ICHINO M, YAGUCHI H. “Generalized Minkowski metrics for mixed feature-type data analysis”. *IEEE Trans Syst Man Cybern 1994, 24:698–707*.

[IRP 2008] IRPINO A, VERDE R. “Dynamic clustering of interval data using a Wasserstein-based distance”. *Pattern Recognition Lett 2008, 29:1648–1658. [Rüs 2001]*

[LEB 84] LEBART L, MORINEAU A, WARWICK KM. *“*Multivariate Descriptive Statistical Analysis”*. New York: Wiley; 1984.*

[ROB 2004] ROBERTSON, S. "Understanding inverse document frequency: On theoretical arguments for IDF". *Journal of Documentation. 60 (5): 503–520. doi:10.1108/00220410410560582*.

[ROY 86] ROYALL RM. (86) “Model robust confidence intervals using maximum likelihood estimators*”. Int Stat Rev 1986, 54:221–226.*

[SU 2016] SU S-F, PEDRYCZ W, HONG T-P, DE CARVALHO AT*. (2016) “*Special issue on granular/symbolic data processing”*. IEEE Trans Cybern 2016, 344–401.*

1. Chapter written by Edwin DIDAY [↑](#footnote-ref-1)