Improving Explanatory Power of Machine Learning in the Symbolic Data Analysis Framework

E. Diday(✉)

CEREMADE, Paris-Dauphine University, Paris, France
diday@ceremade.dauphine.fr

Abstract. Many nice machine learning methods are black box producing very efficient rules but hard to be understandable by the users. The aim of this paper is to help user by tools allowing a better comprehension of these rules. These tools are based on characteristic properties of the original variables in order to remain in the natural language of the user. They are based on three principles, first on local models fitting at best clusters to be found, second on a symbolic description of these clusters and their Symbolic Data Analysis, third on characteristic criterion increasing the explanatory power of the rules by an adaptive process filtering explanatory sub populations.

Keywords: Symbolic data analysis · Machine learning · Symbolic clustering

1 Introduction

In Data Science the aim is to extract new knowledge from Standard, Big and complex data. Another characteristic of Data Science is that its methods and tools are not developed in order to be applied for only a specific domain but for any domain providing data. Often industrial data are unstructured with variables defined on different units. They can also be multi sources (as mixture of numerical and textual data, with images and networks). In order to reduce the size, the complexity and the efficiency of the models associated to such data, a key solution is to use classes, of row statistical units, which are considered as new statistical units. "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. There are at least three advantages of considering classes instead of standard units. First, classes can be the units that interest the most the users. For example, "regions" instead of their "inhabitants" or "species" instead of "specimen", "teams" instead their players or "documents" instead of their "words". Second, classes induce local models often (but not always!) more efficient than global ones. Third, classes give a concise and structured view on the data as they can be organized by a partition, a hierarchical clustering, a pyramid (for overlapping clusters) or a Galois lattice. In clustering classes (called "clusters") are not known and can be obtained by a clustering process called "Dynamical Clustering Method" (i.e. DCM) (see [11]) or for fuzzy classes by EM (see [18]), improving iteratively the fit between each obtained cluster and its local

Y. Hernández Heredia et al. (Eds.): IWAIPR 2018, LNCS 11047, pp. 3–14, 2018.
https://doi.org/10.1007/978-3-030-01132-1_1

associated model. In the case of unsupervised data, clusters can be modeled for example, by means (as in the "k-means" method), distributions (as in mixture decomposition) or factorial axis (which leads to local factorial analysis). In case of supervised data, clusters can be modeled by regressions (or more generally by canonical analysis), neural networks, SVM, etc. In both cases the obtained classes can be described in order to express their within-class variability by vectors of intervals, probability distributions, weighted sequences, functions, and the like, called "symbolic data". Hence, we obtain a symbolic data table that we can study in order to obtain explanatory information on the given classes or obtained clusters. We can also use these symbolic data in order to measure the explanatory power of each class or cluster. More generally 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 Fig. 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 Fig. 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 Fig. 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%

Fig. 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.

More generally, the first characteristic of the so-called “symbolic variable” 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.

The first aim of the so called “Symbolic Data Analysis” (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 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. 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 units are the teams and so, there are classes of individuals.

Complex data constitute an important source of symbolic data. We consider “complex data” as 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 variables coming from multi sources sometimes at multi levels. In this case one of the advantage of “symbolic data” is that unstructured data with unpaired samples at the level of row units, become structured and paired at the classes’ level. By definition, a “class of complex data” is a vector of standard classes defined on different statistical space 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 hospitals, Cs the class of schools, Ci the class of inhabitants, of this region.

Example of complex data, classes and symbolic variables in Official Statistics:

> ***National Statistical Institutes (NSI) organize census in their regions on different kinds of populations: hospitals, schools, inhabitants etc. Each of these populations are associated to their own characteristic 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 of different sizes. If we have n regions and N populations (of hospitals, schools, etc.), then we get after aggregation, a symbolic data table with n rows and p1 + … + pN columns associated to the N sets of symbolic value variables characteristic of each of the N populations. For sure other variables (standard or symbolic) can be added in order to describe other aspects of the regions.***

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

The theory and practice of SDA have been developed in several books [2, 3, 15], many papers (see overviews in [1, 9]), and several international workshops Special issue related to SDA has been published, for example in the RNTI journal, edited by, Guan et al. [20] on 'Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis'; in the ADAC journal on SDA, edited by Brito et al. [4]; in IEEE Trans Cybern [25]. We indicate among many others, four examples of application in nuclear power point [6], epidemiology [21], in cancerology [23], and in face recognition [24].

The paper present three sections after the introduction. The first is devoted to the building of symbolic data from given classes or obtained clusters. The next section shows that the explanatory power of the symbolic data describing a class can be measured by different criteria which provide a measure of the explanatory power of this class. In the next section we show that any tool of machine learning can be transformed by a clustering process in local tools, often more adapted than global ones. Then, based on the explanatory criteria characterizing individuals, classes and variables we show how to improve the explanatory power of any machine learning tool by filtering explanatory sub populations.

2 Building Symbolic Data from Given Classes or Clusters

The aim is to study the symbolic data table provided by the description of given classes (as in supervised learning) or given clusters (obtained from a clustering process), in order to get complementary knowledge enhancing the usual standard interpretation (by means, variance, etc.). For example, in mixture decomposition clustering the description of each class is just given by the analytical expression of the joint probability density $\mathrm{f_i}$ associated to each class. Hence in case of Gaussian model, the joint is described by a big correlation matrix heavy to interpret when there are numerous variables. Building a symbolic data table where the units are the given classes or 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 $\mathrm{L_{k,}}$ is the representative of the class $\mathrm{P_k}$, then the weight $\mathrm{t_k(u)}$ of an individual u in class $\mathrm{P_k}$, which takes the value $\mathrm{x_i^j}$ for the variable j and the individual I, is given by: $\mathrm{t}_k(\mathrm{u}) = d(u, L_k)$ where d is the dissimilarity used by the clustering method which has produced the classes. In case of fuzzy clusters (like in EM), $\mathrm{t}_k(\mathrm{u})$ is the fuzzy weight of u in the kth class. Then, the histogram for the kth class and the jth variable is given by:

$$H_{kj} = \left(\frac{\sum_{i=1}^{N} t_k(x_i)_{\mathrm{I}_1}\left(x_i^j\right)}{\sum_{v=1}^{V} \sum_{i=1}^{N} t_k(x_i)_{\mathrm{I}_v}\left(x_i^j\right)}, \ldots, \frac{\sum_{i=1}^{N} t_k(x_i)_{\mathrm{I}_V}\left(x_i^j\right)}{\sum_{r=1}^{p} \sum_{i=1}^{N} t_k(x_i)\left(x_i^j\right)} \right) \tag{1}$$

where $\delta_{\mathrm{I}_v}\left(x_i^j\right)$ is a vector of Dirac mass defined on V intervals $(I_1, \ldots, I_V)$ partitioning the domain $\mathrm{D_j}$ of the numerical variable $\mathrm{X_j}$ such that: $\delta\left(x_i^j\right) = (\delta_{\mathrm{I}_1}\left(x_i^j\right), \ldots, \delta_{\mathrm{I}_V}\left(x_i^j\right))$ where $\delta_{\mathrm{I}_v}\left(x_i^j\right)$ takes the value 1 if $x_i^j \in \gamma I_v$ and 0 elsewhere. When the I_v are categorical values instead of intervals, we obtain a bar chart and $\delta\left(x_i^j\right)$ takes the value 1 if x_i^j is the category I_v and the value 0 elsewhere.

When instead of a fuzzy partition $(P_1, \ldots, P_K)$ as the one given by EM we have an exact partition denoted $\left(P_1^{'}, \ldots, P_K^{'}\right)$ as the one induced by $P_K^{'} = \{x_i / f(x_i, a_k) \geq f(x_n, a_k)\}$ or directly by DC, we can build in the same way an histogram or a bar chart by setting:$t_k(x_i) = 1$, for any $k = (1, \ldots, K)$ and $i = (1, \ldots, N)$.

In SDA, in order to increase the explanatory power of the obtained symbolic data table, first the chosen number of intervals I_v 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 [16]). After an EM mixture decomposition, the joints $\mathrm{f_i}$ associated to each class $\mathrm{C_i}$ are described by their marginal $\mathrm{f_{ij}}$. These marginal are then described by several kinds of symbolic data as histograms or interquartile intervals or any kind of property, mean, mean square, percentiles, correlation between some characteristic variables and the like. More generally, from any clustering method we obtain a symbolic data table on which SDA can be applied.

3 Explanatory Power of Classes or Clusters from Their Associated Symbolic Data Table

Our aim in SDA is to get a meaningful symbolic data table 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 [3, 7, 8, 15]. In case of histogram value variables an example of discriminating tool is given in [DID 2013] by optimizing the lenght of the histograms intervals. There are at least three ways: distances between rows in each column, to be maximized, entropy in each cell to be minimized, correlations between columns to be maximized. More details are given in [9].

Other kinds of explanatory power of a symbolic data table can be defined. First, we can define a theoretical framework for SDA in the case of categorical variables (see [17, 19]. Let be three random variables C, X, S 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 a categorical variable: Ω → M such that X(w) = x is a category among the set of categories M of this variable. From C and X, we can build a third random variable defined as follows: S: Ω →[0, 1] such that S(w) = s(X(w), C(w)) = s (x, c) the proportion of the category x inside the class c. In other words s(x, c) can be considered as the probability of the category x knowing the class c: s(x, c) = Pr (X = x/C = c). If we denote f $_c$(x) the value of the bar chart induced by the class c and the category x, we have f $_c$(x) = s(x, c).

Characterization of a class by an event: We say basically that a category is "characteristic" of a class if it is frequent in the class and rare in the other classes. In order to insight what we develop in this section, we start with a simple example.

Example: Suppose $f_c(z)$ is higher than 0.9 (i.e. $f_c(x)$ belongs to the event E(z, c)) = [0.9, 1]) and $f_{c'}(x)$ for most of the classes class c' Є P different of c, belongs to the event E(z, c')) = [0, 0.9[). In this case, we can say that the category x is characteristic of the class c versus the other classes of the partition P for the event [0.9, 1] as its frequency takes a value in this event for the class c which is rare for the other classes of the partition P.

A characterization criterion W, varying between 0 and 1, of a category z and a class c can be measured by:

W(z, c, E) = $f_c(z)/(1+g_{z,\ E(z,\ c)}(c))$, where E(z, c) is an event defined by an interval included in [0, 1]containing $f_{\ c}(z)$ and $\Pr(\mathbf{S_x}$ Є E(z, c)) = |{w Є Ω/z = X(w), $\mathbf{S_z}$(w) = s(z, C(w)) Є E(z, c))|/|Ω|, defines g:

$g_{\ z,\ E(z,\ c)}$ (c) = $\Pr(\mathbf{S_z}$ Є E(z, c)). Hence, $g_{\ z,\ E(z,\ c)}$ associates to a class c and a category z, the frequency of individuals of Ω satisfying the event E(z, c) with z = X(w) and c = C(w). If the ground population is infinite, we suppose that Ω is a sample. Hence, given an event E, the criterion W express, how much a category z is characteristic of a class c versus the other classes c' of the given partition P. This criterion means that a category z is even more characteristic 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 z category in any class c' (including c) and satisfying to the event E(z, c) is low in the ground population Ω. Giving z and c, several choices of E can be interesting.

Four examples of events E:

For a characterization of x and c in the neighborhood of s(z, c):

E_1(z, c) = [s(z, c) – ε, s(z, c) + ε] for ε > 0 and s(z, c) Є [ε, 1- ε] where ε can be a percentile.

For a characterization of the higher values than s(z, c): E_2(z, c) = [s(z, c), 1].

For a characterization of the lower values than s(z, c): E_3(z, c) = [0, s(z, c)].

In order to characterize the existence of the category z in any: E_4(z, c) =]0, 1].

Hence, a category z is characteristic of a class c when it is frequent in the class c and rare in a neighborhood of s(z, c) if E = E_1, rare above (resp. under s(z, c) if E = E_2 or E = E_3, rare to appear outside of c in the other classes c' if E = E_4.

In fact there are four cases to consider depending on the fact that in a class c a category z is frequent or not and among the set of classes it is frequent or not in E(z, c). Hence, we have four cases called FF, FR, RF, RR, the cases FF and RR cannot give any specific value to W(z, c, E), but the case FR (resp. RF) where the category is frequent (resp. rare) in c and rare (resp. frequent) in the other classes' leads to a value of W(z, c, E) which is high (resp. low). Therefore, we can say that z is a specific category of c iff W(z, c, E)LogW(z, c, E) is close to 0. Other kinds of characterization criterion can be used. The popular 'test value,' developed in [22], 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.

Characterization of classes and symbolic bar chart variables
Asymbolic data table of bar chart variables can be transformed in a data table where each column is associated to a category, by summing on the characterization of all the cells of each row (resp. column) we obtain a characterization of each class (resp. variable). In the same way, in summing on the characterization of all the cells, we can obtain a characterization of the symbolic data table. In the same way we can find the most typical or atypical class or bar chart variables or symbolic data table. In the following, we focus on characterization but for sure in the same way we could consider the singular or specific case.

It can be shown that the standard Tf-Idf (very popular in text mining) is a case of the W criterion and a parametric version of this criterion can be defined (see [17]).

4 Improving Explanatory Power of Machine Learning by Using a Filter

We show in three steps that any learning machine process can be improved in the efficiency and the explanatory power of its provided rules: in the first step by a dynamical clustering process optimizing at each step a first objective function we obtain local learning models, defined by couples of clusters and local associated predictive models (regression, neural network, SVM, Bayesian, decision tree, etc.) in case of supervised data or couple of cluster and mean, distribution or factorial axis in case unsupervised learning; in the second step the obtained clusters are described by symbolic data (induced by only the explanatory variables in case of supervised data or by all the variables in case of unsupervised data), which leads to the explanatory power of each cluster, measured by a second objective function of characterization; in the third step we provide an allocation rule to any new unit (only known by its explanatory values, i.e. without knowing its predictive values, in case of supervised data) if it improves simultaneously the first and the second objective function (i.e. at least improving one without degrading the other). Several kinds of allocation rules are proposed including Latent Dirichlet models (see Diday [19]).

Hence, in the first step, we use **The "Dynamical Clustering Method" (DCM):** Starting from a given partition $P = (P_1, \ldots, P_k)$ 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 a point x of the population: $f(x) = C$ in order to improve a given criterion at each step until convergence.

Proof. starting from a partition $P = (P_1, \ldots, P_K)$ of the initial population, the representation function applied to the classes P_i produces a vector of representation $L = (L_1, \ldots, L_K)$ among a given set of possible representations, where $g(C_i) = L_i$. A quality criterion can be defined in the following way: $W(P, L) = \sum_{i=1}^{k} f(P_i, L_i)$ where w measures the fit between each class P_i and its representation L_i it decreases when this fit increases.

Starting from a partition $P^{(n)}$, the value of the sequence $u_n = W(P^{(n)}, L^{(n)})$ decreases at each step n of the algorithm. Indeed, during a the allocation step an individual x belonging to a class $P_i^{(n)}$ is affected to a new class $P_j^{(n+1)}$ iff $W(P^{(n+1)}, L^{(n)}) \leq W(P^{(n)}, L^{(n)}) = u_n$. Then, starting from the new partition $P_j^{(n+1)}$, we can always define a new representation vector $L^{(n+1)} = \left(L_1^{(n+1)}, \ldots, L_K^{(n+1)}\right)$ where for any i = 1 to K, $L_i^{(n+1)} = g\left(P_i^{(n+1)}\right)$ fit best to $P_i^{(n+1)}$ than $L_i^{(n)}$ or remains unchanged (i.e. $L^{(n+1)} = L^{(n)}$. This means: $f\left(P_i^{(n+1)}, L_i^{(n+1)}\right) \leq f\left(P_i^{(n+1)}, L_i^{(n)}\right)$ for i = 1 to K.

Hence, at this step, we have $u_{n+1} = W(P^{(n+1)}, L^{(n+1)}) \leq W(P^{(n+1)}, L^{(n)}) \leq W(P^{(n)}, L^{(n)}) = u_n$. As this inequality is true for any n, this positive sequence decreases and converges.

Moreover, notice that in the case where $W(P_i, L_i) = \sum_{w \in P_i} f(w, L_i)$, the allocation step consists to change w from one class to another when $f(w, L_j) < f(w, L_i)$. Notice also that a simple condition of convergence is that for any C_i taken among all the possible subsets of the given population and L_i taken among the given set of possible representations: $f(C_i, g(C_i)) \leq f(C_i, L_i)$.

In case of unsupervised data, the classical k-means method is the case where L_i is the mean of the class C_i. When L_k are probability densities, we have a mixture decomposition method 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. There are many other possibilities such as when representation of any class can be a distance, a functional curve, points of the population, a factorial axis etc. For an overview see [12, 13].

In case of supervised data we can settle for example: $W(P_i, L_i) = \sum_{w \in P_i} f(w, L_i)$ and f (w, L_i) = $\|Y(w) - M_i(w)\|$ where Y(w) is the predictive value given by the supervised data sample and $M_i(w)$ is the value given by the model M_i applied to the class P_i. The convergence of the method is then obtained if for any C_i and L_i taken among a given family of models, $f(C_i, g(C_i)) \leq f(C_i, L_i)$ where $g(C_i) = M_i$ is the best fitting model to the class C_i among a given family of models.

For example in the representation by a regression, each individual is allocated to the class C'_i if this individual fit the best the regression L_i among all the possible regressions, (see [5]), more generally in case of representation by canonical axis see [14]. Notice that this method contains in case of unsupervised data: local PCA (Principal Component Analysis) (See Fig. 2) and local correspondence analysis. In case of supervised data it contains local regression (see Fig. 2) and local discriminant analysis (See Fig. 2).

Notice that we can extend this fuzzy partitioning method in order to get fuzzy local models by considering the $f_i(x, a_i)$ to be the fit between x and a model M_i with parameters a_i.

In the second step we enhance the explanatory power of the clustering by a characterization measure. The characterization measure of an individual w for the jth

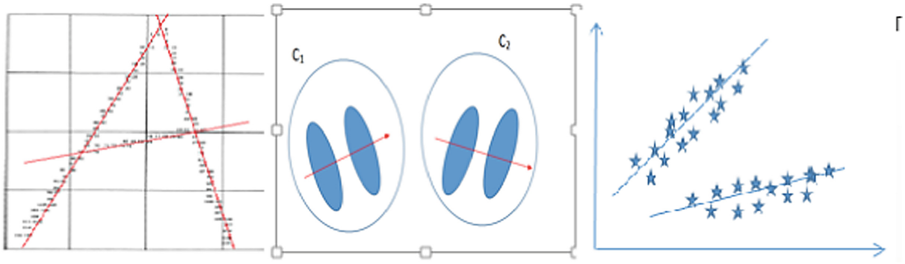

Fig. 2. Local PCA: find simultaneously classes and first axes of local PCA which fit the best Local Discriminant Analysis: find simultaneously classes and first axes of local factorial discriminant analysis which fit the best. Local Regression: find simultaneously classes and local Regressions which fit the best.

variable, the kth class and the event E with: $x = (x_1, \ldots, x_p)$, $x_j = (x_{j1}, \ldots, x_{jp})$, $X_j(w) = x_{jm}$, $C(w) = c_k$ is defined by: $\mathbf{W}(x_{jm}, c_k, E) = f_{c}(x_j)/g_{x,\, E(xjm,\, ck)}(c_k)$ Therefore, we can define a characterization measure of an individual $CI(w) = \sum_{j=1,\, p} W(x_{jm}, c, E)$.

We can define a characterization measure of a symbolic variable X_j by:

$CV(X_j) = \sum_{k=1,\, K} \text{Max}_{m=1,\, mj} W(x_{jm}, c_k, E)$. We can also define a characterization measure of a class c by:

$CC(c) = \sum_{j=1,\, p} \text{Max}_{m=1,\, mj} W(x_{jm}, c, E)$.

We can then place in order from the less to the more characteristic the individuals w, the symbolic variables X_j and the class c_i by using respectively the CI, CV or the CL characteristic measure. All these criteria can then enhance the explanatory power of the local machine learning tool used. These orders are respectively denoted O_{CI}, O_{CV}, O_{CL}.

In the third step we suppose that we have already obtained a clustering from a basic sample where the predictive values are given in case of supervised data. Then, the aim is to allocate new individuals to their best cluster. 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. If the representative is a distribution like in Mixture decomposition 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, an individual can place the clusters in an order denoted O_1.

In case of supervised data the aim is first to allocate a new individual (which predictive value is not given) to the best cluster and then to obtain its predicted value from the local model associated to this cluster. For example, if we allocate a new individual to a cluster modeled by a local regression, 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 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 then use, on these data, a supervised machine learning tool for which any individual can have an order of preference to the clusters from the best allocation to the worse. Hence, by this way, an individual can place the clusters in an order denoted O_2.

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 f_j, we can use the likelihood product of the $f_j(x_j)$ for j = 1, p where x_j 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 $f_j(x_j)$ by W(xj, c, E) in the categorical case. Hence, by this way, an individual can place the clusters in an order denoted O_E. Finally given a new unit, we can place in order the obtained clusters by two ways: O_E or O_i (i = 1 or 2) where O_E is an explanatory order. Several strategy 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 local models.

Machine learning filtering strategies:

The idea is to add (i.e.to filter) a new individual to the cluster and to a symbolic description if it improves simultaneously at best the fit between the cluster and its representative (i.e. its associated model in case of supervised learning) and the explanatory power of its associated symbolic description.

The first kind of filtering strategy is to continue the learning process only with only the individuals which have at best position the same cluster in the order O_E and O_i. Another kind of strategy is to continue the learning machine process with only the individuals whose clusters at best position are not more fare then a given rank k. Then the individual is allocated to the cluster of best rank following O_E or O_i alternatively or depending if you wish more explanatory power or better decision. Other strategies are also possible by adding O_{CI}, and (or) O_{CL} to O_E and O_i. It is also possible to reduce the number of variables by choosing the first ones in the O_{CV} order. 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 lead to other subpopulations.

5 Conclusion

We have first introduced Symbolic Data Analysis which can give useful complementary knowledge to any standard data analysis. We have recalled local data analysis obtained by Dynamic Clustering which can give more accurate results to any kind of data analysis. We have defined several kinds of characterization criteria which allow to place in order individuals, clusters and variables following their explanatory power. We finally gave several strategies for filtering individuals which give the best explanatory of the machine learning process by alternatively improve rules and explanatory power.

Much remains to be done in order to compare and improve the different criteria and strategies and to test the results with different black box machine learning methods (Neural network, SVM, Deep machine learning, etc.) on different kinds of data.

References

1. Billard, L., Diday, E.: From the statistics of data to the statistic of knowledge: symbolic data analysis. JASA J. Am. Stat. Assoc. **98**(462), 470–487 (2003)
2. Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley Series in Computational Statistics, p. 321. Wiley, Chichester (2006). ISBN: 0-470-09016-2
3. Bock, H., Diday, E.: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, p. 425. Springer, Heidelberg (2000). https://doi.org/10.1007/978-3-642-57155-8. ISBN: 3-540-66619-2
4. Brito, P., Noirhomme-Fraiture, M., Arroyo, J.: Special issue on symbolic data analysis. Adv. Data Anal. Classif. **9**, 1–4 (2015)
5. Charles, C.: Régression typologique et reconnaissance des formes. Thèse de 3ème cycle, Juin 1977, Université Paris IX-Dauphine and INRIA Rocquencourt 78150 (France) (1977)
6. Courtois, A., Genest, Y., Afonso, A., Diday, E., Orcesi, A.: In service inspection of reinforced concrete cooling towers – EDF's feedback, IALCEE 2012 Vienne. Autriche (2012)
7. De Carvalho, F.A.T.: Extension based proximity coefficients between constrained Boolean symbolic objects. In: Hayashi, C., et al. (eds.) Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 370–378. Springer, Berlin (1998). https://doi.org/10.1007/978-4-431-65950-1_41
8. De Carvalho, F., Souza, R., Chavent, M., Lechevallier, Y.: Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recogn. Lett. **27**, 167–179 (2006)
9. Diday, E.: Thinking by classes in data science: symbolic data analysis. In: WIREs Computational Statistics Symbolic Data Analysis, vol. 8, p. 191, September/October 2016 © 2016 Wiley Periodicals, Inc. (2016)
10. Diday, E.: The Dynamic clusters method in non-hierarchical clustering. Int. J. Comput. Inf. Sci. **2**(1), (1973). https://doi.org/10.1007/bf00987153
11. Diday, E., Schroeder, A.: A new approach in mixed distributions detection. RAIRO **10**(6), 75–106 (1975)
12. Diday, E., Simon, J.C.: Clustering analysis. In: Fu, K.S. (ed.) Communication and Cybernetics Digital Pattern Recognition, vol. 10, pp. 47–94. Springer, Berlin (1979). https://doi.org/10.1007/978-3-642-67740-3_3
13. Diday, E., et al.: Optimisation en classification automatique, INRIA publisher (2 books 887 pages). INRIA, 78150 Rocquencourt, France (1980). ISBN 2-7261-0219-0
14. Diday, E: Canonical analysis from the automatic classification point of view. Control Cybern. **15**(2) (1986)
15. Diday, E., Noirhomme-Fraiture, M. (eds.): Symbolic Data Analysis and the SODAS software. Wiley, Chichester (2008). ISBN 978-0-470-01883-5
16. Diday, E., Afonso, F., Haddad, R.: The symbolic data analysis paradigm, discriminate discretization and financial application. In: Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis, HDSDA 2013. Revue des Nouvelles Technologies de l'Information vol. RNTI-E-25, pp. 1–14 (2013)

17. Diday, E.: Explanatory power of clusters based on their symbolic description. In: Saporta, G., Wang, H., Diday, E., Guan, R. (eds.) Advances in Data Sciences. ISTE-Wiley (2019)
18. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data with the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. **39**, 1–38 (1977)
19. Emilion, R., Diday, E.: Symbolic data analysis basic theory. In: Saporta, G., Wang, H., Diday, E., Guan, R. (eds.) Advances in Data Sciences. ISTE-Wiley (2019)
20. 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)
21. Guinot, C., Malvy, D., Schemann, J.-F., Afonso, F., Haddad, R., Diday, E.: Strategies evaluation in environmental conditions by symbolic data analysis: application in medicine and epidemiology to trachoma. ADAC (Adv. Data Anal. Classif.) **9**(1), 107–119 (2015)
22. Lebart, L., Morineau, A., Km, W.: Multivariate Descriptive Statistical Analysis. Wiley, New York (1984)
23. Nuemi, G., et al.: Classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy. Cancer Epidemiol. J. **37**, 688–696 (2013)
24. Ochs, M., Diday, E., Afonso, F.: From the symbolic analysis of virtual faces to a smiles machine. IEEE Trans. Cybern. **46**(2), 401–409 (2016). https://doi.org/10.1109/tcyb.2015.2411432
25. Su, S.-F., Pedrycz, W., Hong, T.-P., De Carvalho, F.A.T.: Special issue on granular/symbolic data processing. IEEE Trans. Cybern. 344–401 (2016)