

# The symbolic data analysis paradigm, discriminant discretization and financial application

Edwin Diday\*\*, Filipe Afonso\*, Raja Haddad\*, \*\*\*

\*Syrokko, Aéroport Roissy CDG, 95731 Roissy  
(haddad, afonso)@syrokko.com

(\*\*CEREMADE, \*\*\*LAMSADE), Université de Paris 9 Dauphine, 75775 Paris  
diday@ceremade.dauphine.fr

**Abstract.** The variability inside classes of individuals, categories (defined by a categorical variable) or concepts (defined by an intent and an extent, like species for example), is expressed by the use of intervals, histograms, distributions, sequences of weighted values and the like. In this way we obtain new kinds of data called "symbolic". The aim of "Symbolic Data Analysis" (SDA) is to study and extract new knowledge from these new kinds of data by an extension of Statistics and Data Mining to symbolic data. We show that SDA is a new paradigm opened to a vast field of research and applications. Then, we give a way for obtaining discriminate symbolic descriptions by an original discretisation method, which is illustrated by a financial application.

## 1 Introduction

The usual data mining model is based on two parts: the first concerns the observations (i.e., observed entities), the second contains their description by several standard numerical or categorical variables. The Symbolic Data Analysis (SDA) model (see (Diday, 1987), (Billard and Diday, 2006), (Diday and Noirhomme, 2008)) needs two more parts: the first concerns higher level units defined by classes of observations, categories (i.e. a name given to a class) or concepts (defined by an intent and an extent) and the second concerns their description by "symbolic data" which may be standard categorical or numerical data but moreover intervals, histograms, sequences of weighted values and the like, in order to take care of the variability of the observations inside each class. These new kinds of data are called "symbolic" as they cannot all be manipulated like numbers.

Based on this model, new knowledge can be extracted by new tools of data mining extended to this higher level units considered as new kinds of observations. Inspired by "The Structure of Scientific Revolutions" (Kuhn, 1962), the second section of this paper tries to show that the "Symbolic Data Analysis" framework is a new scientific paradigm by answering the following questions: what is the failure in the actual practice? What is the paradigm shift? What is to be observed and scrutinized? What kind of questions are there, and how are they structured? What are the principles and the theoretical development? What is the applicability domain?

In order to build symbolic data from an initial standard data table, a discretization process is needed and constitutes a fundamental part of the aggregation process which leads to histogram

The symbolic data analysis paradigm

valued variables. Discretization is the process of converting numerical values of an attribute to categorical. Therefore, in the third section we present a method, called "HistSyr", which converts a continuous attribute to histograms. The main purpose of this method is to discover the discretization of the continuous variable that will provide the most different histograms, from one higher level unit (class or category) to another, after the aggregation to symbolic data. In the last section, the paper is illustrated by an application on financial data for the classification of investment funds (described by symbolic data), as a decision support.

## **2 The SDA paradigm**

### **2.1 What is the actual failure which has produced the SDA Paradigm?**

The failure is that in the actual practice only the "individual" kind of observations is considered. Moreover, these individual observations are only described by standard numerical and categorical variables.

What is the SDA paradigm shift? It is the transition from the analysis of "individual observations" (e.g. a player of soccer, a stock, a pig etc..) described by standard variables of numerical and categorical values, to the analysis of "classes of individual observations" (a team of players, a fund of stocks, a farm of pigs, etc..) considered as higher level observations described by variables whose values take into account the variability inside the classes: intervals, probability distributions, sets of categories or numbers, random variables, etc..

These new kinds of values are called "symbolic" and the variables of symbolic values are called "symbolic variables". For example, if we want to know what makes a team win, the units to be considered are the teams and not the players. Moreover the description of a team needs the use of symbolic data since for example, the age of a player is 24 years old, the age associated to a team of players is for example the [min, max] interval or a histogram of the ages of the players of the team. More generally, this transition is needed in the case of "complex" data as it will be explained hereunder.

### **2.2 What is to be observed and scrutinized?**

First, standard data tables where standard numerical and (or) categorical variables induce categories which can be considered as "higher level observations". These categories can be described by symbolic variables taking into account of their internal variability. The obtained symbolic variables are more or less discriminating these categories depending on the discretization process (section 4), used on the variables of the initial standard data table.

Second, "native symbolic data" represented by tables where the observations are of higher level and the ground level of observations is not known (for example, species of birds or species of mushroom are higher level observations, described by their interval of height, age, distribution of colors etc., well defined in specialized books where the ground level of the specimens is not given).

Third, "Complex" data where several different sets of individuals are described by different variables. The next section describes this case into more detail.

### 2.3 From complex data to symbolic data table: the fusion process

"Complex" data cover all sets of data which cannot be reduced to a simple standard data table. For example, hierarchical data, multisource data, specific data like text or images. In practice, often complex data are more or less based on several kinds of observations described by standard numerical or (and) categorical variables contained in several multisource related data tables. In this case the higher level units are obtained by means of a "fusion process".

For example, in the study of the degradation problems occurring on nuclear power plant cooling towers, we can consider, for each tower, three standard data tables of different observations and different variables. The observations of the first data table are cracks described by their length, thickness orientation etc. In the second standard data table the observations are vertices of a grid obtained on the tower and described by the gap deviation at different periods compared to the initial model position. In the third standard data table, corrosion positions are described by corrosion variables. In order to compare the towers, a fusion process is applied. It consists first of an aggregation process on the three data tables of each tower considered as a higher level observation and then of a concatenation of the obtained aggregated tables. Tab-Syr software is used to create, fusion and visualize symbolic data table. Figure 1 illustrates an example of graphical illustration of symbolic data table obtained after the fusion process (see (Afonso et al., 2010) for more details).

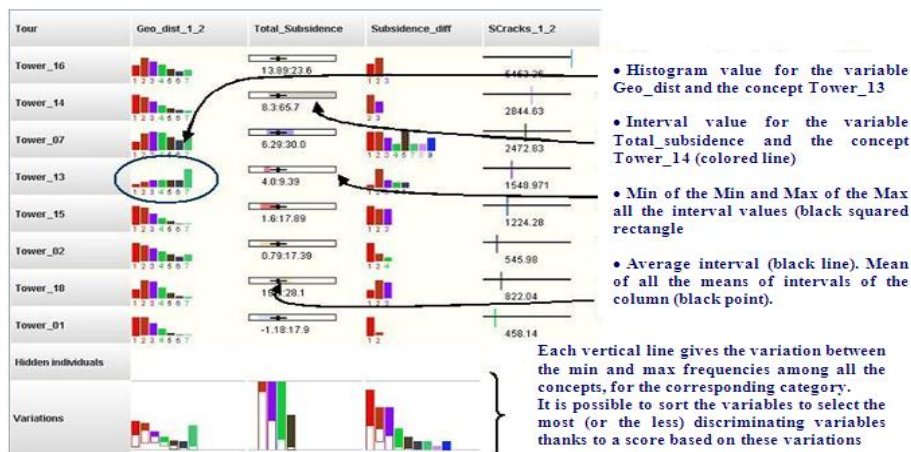


FIG. 1 – Visualization of the symbolic data table (rows are sorted from the most to the less damaged towers thanks to a model that combines different kinds of symbolic variables such as histograms, intervals, standard variables...).

### 2.4 What are the main principles?

#### 2.4.1 The general methodology: four approaches

The methodology can be decomposed into four cases, depending on the initial data (standard or symbolic) and the used method (standard or symbolic). The observed entities can be

## The symbolic data analysis paradigm

natively described by symbolic data, for example, after a fusion process as shown in the section 2.3. If the initial data are standard, for sure standard methods may be applied for describing individuals (for example, allocating individuals to the labelled leaves of a decision tree). Nevertheless, it is also possible to transform these standard data using clusters (after a clustering process) or categories (given or obtained after a discretization process). Then, these clusters or categories are considered as new units to obtain symbolic data (by an aggregation process). Then, an SDA method in order to solve the same question can be used (for example, allocating individuals to the labelled leaves of a symbolic decision tree). For example, having players described by standard numerical and categorical data, if the question is to find what explains that a player is good, a standard decision tree method can be applied. Nevertheless, we can also solve the problem after the transformation of the initial data in categories or clusters, considered as new units, described by symbolic data and then apply a Symbolic Decision Tree method.

If the initial data are natively symbolic, we can solve the problem by also two approaches: first, the standard approach which transforms the symbolic data in standard data and then applies a standard method; second, by using an SDA method on the symbolic data themselves. For example, if the question is "what makes a team win?", the initial data are symbolic in order to describe the variability inside the teams considered as units. Then, a standard or an SDA method can be applied on these new units.

### 2.4.2 Symbolic data express variability

There are at least two kinds of uncertainty: "imprecision", "vagueness". Both concern individuals value variables. "Variability" concerns individuals value variables variation inside a class of individuals. For example, Zidane (the famous football soccer player) has a height of 1.80 cm with 1 cm of "imprecision", we can say "vaguely" that "frequently" he shoots goals, his age was precisely 32 years old but the age of his team at the football World Cup "varies" between 23 (exact age of the younger player of his team) and 34 (exact age of the older player). More generally, this means that the variability can be expressed with exact information (i.e. without imprecision or vagueness). For sure, it is possible to have measures mixing the three notions as for example: the variation of the age of the players of Zidane's team where the ages of the players are known with imprecision or (and) vagueness.

### 2.4.3 The SDA basic principles

When SDA is applied to numerous standard data in order to solve a standard decision problem, in many cases, in comparison with a standard method, the SDA approach has the advantages to produce simpler and efficient results, easier to explain than those obtained by a standard method. From the ground level (individual observations) to a higher level of observation (classes, categories or concepts), an aggregation process which produces the symbolic data is needed. The aggregation process must discriminate as much as possible the Symbolic Data obtained at the higher level observations by using a "good" discretization process of the initial variables. When SDA is applied to symbolic data, in comparison with the transformation of the symbolic data in standard data, the advantages are that:

- The variables taking symbolic values are not lost. For example, if a team is described by a vector of two intervals expressing the height and the age of its players, in the standard

space we are in the  $\mathbb{R}^4$  space where each team is represented by a point associated to four numerical variables (min and max of the height and min and max of the age), but in the symbolic space we have only two variables the symbolic variables height and age of interval values easily representable by a rectangle of  $\mathbb{R}^2$ . Therefore, a standard decision tree applied in  $\mathbb{R}^4$  will concern the variables min or max of height or age and will lose the variables height and age as a symbolic decision tree (see (Quantin et al., 2011)) applied in  $\mathbb{R}^2$  will do.

- The standard space fails to represent graphically the variability. For example, a standard PCA (see (Jolliffe, 2002)) applied in  $\mathbb{R}^4$  will represent the teams by points. At the contrary in a Symbolic PCA (see (Douzal-Chouakria et al., 2011)), the teams are represented by a surface better expressing the variability than just a point.

So, in summary we can say:

- Work in the Symbolic space (without transforming symbolic data into standard data) as much as possible. The symbolic data has to be treated as they are as their transformation into standard numerical or categorical data, loose the symbolic variables and the variability that they express. For more details see (Billard and Diday, 2006) (pages 63 to 65).
- Represent as much as possible, the graphical visualization of the variability of the higher level observations in the output.
- Do not reduce the output to just numerical results and try to remain in the same or more general kind of symbolic data than in the input.

## 2.5 What is the theoretical development?

Like for standard Statistics or standard Data Mining there is not a general theory but specific problems to solve. All needed theoretical development in Statistics or Data Mining for analyzing standard data can be extended to symbolic data. Moreover, specific theoretical developments are needed. For example, in the case where the ground data are given and the histogram valued variables are built, after a "well discriminating" (see section 3) discretization process, the stochastic convergence of any SDA method has to be proved when the histograms converge towards their associated laws when they exist (see for example, (Diday and Emilion, 2003) in the framework of Galois lattice theory). More generally, there are three kinds of theoretical developments depending on the input (standard or symbolic data), and the output (standard or symbolic). For example, what links, advantages or inconvenients are there in applying a standard method (PCA, decision tree, clustering, mixture decomposition, etc.), on standard data and using the associated symbolic method by transforming these data to symbolic data? This strategy can be at least useful in the case of Big Data. What links between a parametric model induced from symbolic data transformed in standard data and a parametric model induced from the model associated to each of these symbolic data (see for example: (Le-Rademacher and Billard, 2011), (Diday, 2011))? The question of aggregating standard data in order to obtain discriminating symbolic data (see section 3, the case of histograms) remains open. The question of aggregating (by capacities, for example) symbolic data and crossing (by copulas, for example) symbolic variables also remains wide open.

The symbolic data analysis paradigm

## 2.6 What is the SDA applicability domain?

The applicability domain is huge, as it is the case for Data Mining and any domain where multidimensional data are massively collected: medicine, biology, socio-demography, economics, text mining, images, web data, social networks, etc.

## 3 Discretization of continuous variable

In this section we present the state of the art of a continuous variable discretization. Then we explain the notion of discrimination. After this, we present the "HistSyr" software. Finally, we give a comparison between HistSyr and other discretization methods.

In the literature, we find many works which treat discretization problematic especially to resolve the problem of the data mining methods that accept only categorical variables in input like decision tree and naïve-Bayes (see (Yang and Webb, 2001)).

There are two types of discretization ((Dougherty et al., 1995)):

- Unsupervised discretization in which we do not take into account the value of the classes of the objects, like Equal Width Discretization (EWD), Equal Frequency Discretization (EFD) and the Fisher's algorithm (Fisher, 1958) which gives the best bounds by optimizing the intra-class inertia of the continuous variable.
- Supervised discretization in which the value of the classes of the objects is essential to choose the best cutting bounds. In every method of this family there is a criterion of discretization which is calculated on the values of the classes. For example the multi interval algorithm (MIA) of Fayyad and Irani ((Fayyad and Irani, 1993)) this criterion is Shannon Entropy, while it is  $\chi^2$  statistic in the ChiMerge algorithm of Kerber (Kerber, 1992).

### 3.1 What's discrimination?

In this context, the meaning of discrimination is differentiation between concepts. In fact the more the values of an attribute are different the more this attribute is discriminating the concepts. Figure 2 gives an example of discriminating and not discriminating variables. Based on this example, we note that the first variable has values that differentiate perfectly the concepts. How can we obtain discriminating histograms from continuous variables? To answer this question, we have developed a software module called HISTSYR.

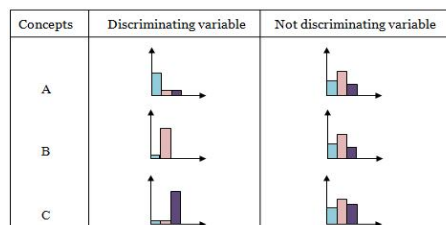


FIG. 2 – Discriminating vs. not discriminating variables.

### 3.2 HistSyr

The main purpose of HistSyr is to build discriminating histograms from continuous attributes. Based on the principle of classical supervised discretization methods we construct a discretization criterion which takes into account the different values of the concept and finds the bounds that give the most discriminating histograms for the given concepts. This criterion is based on calculating the Euclidean distance between two histograms  $H_i$  and  $H_j$ . The expression of this criterion is:

$$Score = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{j=1}^k abs(mod(i, j) - mod(l, j)) \quad (1)$$

where  $n$  is the number of concepts,  $k$  is the number of categories and  $mod(i, j)$  represents the frequency of the category  $j$  for the concept  $i$ .

This criterion evaluates the difference between the constructed histograms. The results are better as the score is higher. As shown in figure 3 the value of the criterion varies between 0 and 1. It is equal to "1" when the histograms are perfectly different and "0" when they are identical.







	Case 1	Case 2	Case 3
Concept A			
Concept B			
Score	1	0	$0 < score < 1$

FIG. 3 – Examples of calculating the score.

To search for the best bounds which give the best histograms we use Fisher's algorithm ((Fisher, 1958) ), except that instead of optimizing the internal variation of the continuous variable we optimize the score presented previously (see equation 1). So we search the bounds which give the best score. Using Fisher's algorithm to search bounds we make sure that we obtain the optimal solution, since this algorithm searches the best solution from all possible bounds.

### 3.3 HistSyr Vs other methods

To validate our method it was important to compare it to other methods of discretization. To do this we have chosen methods that discretize a continuous variable into several intervals like EWD, EFD and the method of Fayyad an Irani (see section 3.2). We have implemented the two first methods in HistSyr and the result of the last method was found using the software Weka Ware (2000). To compare these methods we test them on several machine learning data sets from UCI (Blake and Merz, 1998)).

## The symbolic data analysis paradigm

Figure 4 gives an example of results obtained by HistSyr applied to the Iris database. In this example we take the continuous variable "SepalLength" and we construct histograms resulting from the four methods ; the first column gives the result of the application of "HistSyr" with 3 categories ; the second column gives the result of the EWD with 3 categories, while the third represents the results of EFD method with 3 categorie ; finally the fourth coloum gives the result of the application of Fayyad and Irani's method. This example shows that the supervised methods, "MIA" and "HistSyr", produce more discriminating histograms than the unsupervised ones.

Species	SepalLength_Hist3	SepalLength_inter...	SepalLength_Area3	SepalLength_limit3
setosa				
versicolor				
virginica				
Scores	<b>0.71</b>	<b>0.66</b>	<b>0.65</b>	<b>0.70</b>

FIG. 4 – Example of results obtained by the application of four discretization methods to the attribute "Sepal Length" of the Iris database.

Table 1 resumes the obtained results of these methods with three UCI databases: "Iris", "Breast cancer Wisconsin" and "Australian".

This table shows that our method performs better according to our criteria than the unsupervised methods (EWD and EFD). But it gives in some cases the same scores as the method of Fayad and Irani with a less number of categories like for the attribute "Sing\_epit". We may conclude that HistSyr lead to the best bounds for histograms that discriminate concepts.



Databases	Attributes	HistSyr		EWD		EFD		MIA	
		Bounds	Score	Bounds	Score	Bounds	Score	Bounds	Score
Iris	Sepallength	5.45; 6.15	0.71	5.55; 6.7	0.66	5.4; 6.3	0.65	5.55; 6.15	0.70
	Sepalwidth	2.95; 3.05	0.48	2.8; 3.6	0.29	2.9; 3.2	0.44	2.95; 3.35	0.47
	Petallength	2.45; 4.85	0.95	2.97; 4.93	0.94	2.45; 4.9	0.95	2.45; 4.75	0.95
	Petalwidth	0.8; 1.65	0.96	0.9; 1.7	0.96	0.8; 1.6	0.95	0.8; 1.75	0.96
Cancer	Clump	4.5	0.64	4; 7	0.60	3; 5	0.64	4.5; 6.5	0.64
	Uni_cel_size	2.5	0.86	3.25; 5.5; 7.75	0.82	1; 5	0.71	1.5; 2.5; 4.5	0.86
	Uni_cel_sha	2.5	0.85	3.25; 5.5; 7.75	0.82	1; 5	0.72	1.5; 2.5; 4.5	0.85
	Marg_adh	1.5	0.69	4; 7	0.64	1; 3	0.68	1.5; 3.5	0.69
	Sing_epit	2.5	0.81	4; 7	0.69	2; 3	0.81	2.5; 3.5	0.81
	Blan_chrom	3.5	0.77	4; 7	0.77	2; 3	0.64	2.5; 3.5	0.77
	Norm_nucl	2.5	0.75	4; 7	0.65	1; 2	0.71	2.5; 8.5	0.75
	Mitoze	1.5	0.42	5.5	0.13	1.5	0.42	1.5	0.42
Australian	A2	36.625	0.16	47	0.09	28.625	0.08	38.96	0.15
	A3	4.208	0.23	14	0.05	2.75	0.20	4.2075	0.23
	A7	1.02	0.39	14.25	0.04	1	0.36	1.02	0.39
	A10	0.5	0.46	22.33; 44.66	0.01	0.5; 9.5	0.46	0.5; 2.5	0.46
	A13	105	0.22	1000	0.01	160	0.15	99.5	0.21
	A14	232	0.35	50000	0.01	6	0.22	493	0.33

TAB. 1 – Results of applying HistSyr, EFD, EFW and MIA on 4 databases

The symbolic data analysis paradigm

## 4 Application: Clustering of investment funds

This section describes an application of symbolic data methods to the classification of investment funds as a decision support for the establishment of financial portfolios.

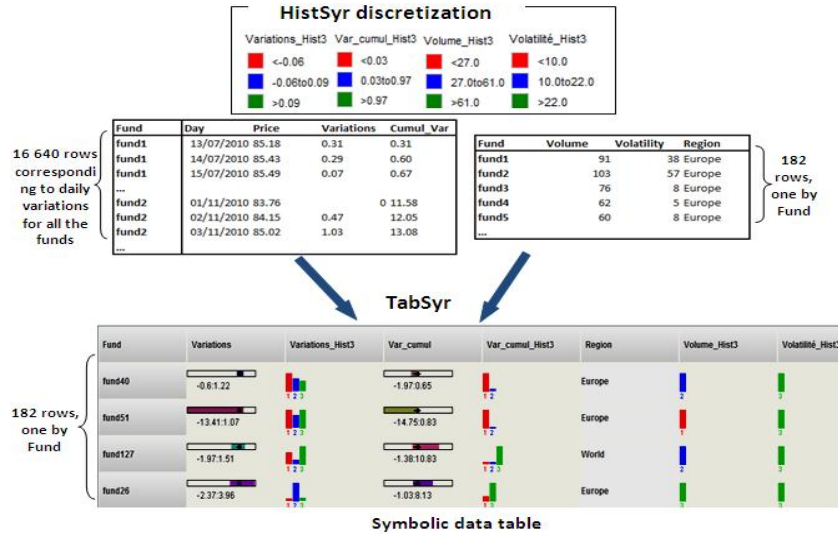


FIG. 5 – Discretization by HISTSYR and then merging of classical data tables in a single symbolic data table describing the investment funds. The variables describing the funds can be interval or histogram valued in order to take into account the daily variations of the funds.

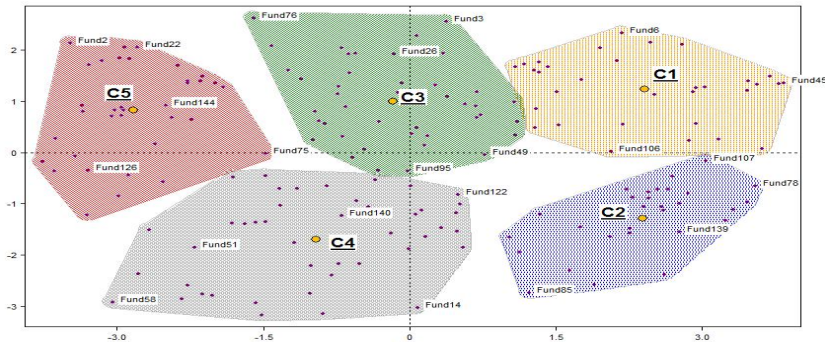


FIG. 6 – Principal Component Analysis (PCA) of investment funds and projection of a k-means clustering of these funds into five clusters (C1 to C5). The PCA and k-means methods are here extended to symbolic data.

Data refer to 182 non-denominated investment funds described by two data files. The first data file describes the daily variation of funds during three months. Each row of the file collects the fund, the trading day, the price in Euros of the fund, the variation from the previous day and the cumulative variation (Cumul\_ Var) from time 0 (first day of the quarter). This file contains

16,640 rows, one row by trading day for each fund. The second file gives, for each fund, the volume of funds traded in the quarter (in millions), the daily volatility (i.e. the variance of the prices for the period) for the quarter (100) and the original region of the fund. This file contains 182 rows, one by fund. An extract of the tables is given at the top of Figure 5.

To form portfolios of investment funds, the concept "investment funds" was built. Thus, a symbolic data file is built using the module TabSyr for fusion. Tabsyr is applied to the first initial data file on variations. With this fusion, the values of the variables "Price", "Variations" and "Cumul\_Var" are aggregated up to intervals describing the min/max variations of a giving fund over the whole period (see bottom of figure 5). Moreover, in order to take into account the daily variations in a more informative way, the variables "variations" and "Cumul\_Var" are aggregated up into histograms. To perform this aggregation, Histsyr methodology is used to ensure the definition of relevant histograms, i.e. histograms that discriminate the "investment fund". "variations\_hist3" describes, for a given fund, the daily variations over the whole quarter in the classes ( $<-0.06$ ,  $-0.06$  to  $0.09$ ,  $>0.09$ ). Thus, for a given histogram, the red bar, in the symbolic data table, expresses the frequency of the number of days were to perform less than  $-0.06$ , the blue bar expresses frequencies of returns between  $-0.06$  and  $0.09$ , and finally the green bar expresses frequencies of daily returns over  $0.09$ . To make it simpler, we may consider that each histogram expresses the loosing, steady and winning days respectively. Similarly, the "Var\_cumul\_hist3" variable describes the cumulative variations thanks to histograms distributed in the classes ( $<0.03$ ,  $0.03$  to  $0.97$ ,  $>0.97$ ).

In a second step, the second file on the individual variables is concatenated to the symbolic data file. The Histsyr method is used again to discretize the quantitative variables "Volume" and "Volatility" into categories. The three variables "Region", "Volume\_hist3" ( $<27$ ,  $27$  to  $61$ ,  $>61$ ) and "Volatility\_Hist3" ( $<10.0$ ,  $10.0$  to  $22.0$ ,  $>22.0$ ) are added to the symbolic data file. Finally, a symbolic data table merging all variables was created where each cell of the table does not necessarily contain a single value but interval or histogram values taking into account the daily variations of the funds. This merging process is illustrated in figure 5.

The resulting table is then analyzed thanks to data analysis methods extended to symbolic data and implemented in SYR software. Using the NetSyr module, extending, among other options, Principal Component Analysis (PCA) to symbolic data (see (Diday, 2010)), the investment funds are visualized in a factorial plane (see figure 6). The first axis captured 36% of the inertia and the second one, 18%. In this factorial plane, a clustering of the funds into 5 clusters, calculated with the k-means method for symbolic data (see chapter 11 of De Carvalho and al. in (Diday and Noirhomme, 2008)), is projected through colored shapes. These five clusters of funds, numbered from C1 to C5, will be analyzed in the following sections.

To perform this analysis, several visualization tools offered by the NetSyr module were used. In Figure 7, the quarterly volatility of each fund combined with its region is visualized using colored rings. The larger the circle, the higher the volatility for a given fund. The colors in each ring correspond to the regions: Europe (Blue), U.S. (green), World (Red). We find that the most volatile funds appear mostly on the far left of the graph. We also note that the five clusters of funds are mixed considering the regions. We can, however, observe, in some clusters, several small groups (subclusters) of very similar funds originated from the same region. The projection of a proximity network between funds can bring out these small groups. Examples were encircled in the graph. Finally, the proximity network allows observing funds on the frontier between two clusters.

The symbolic data analysis paradigm

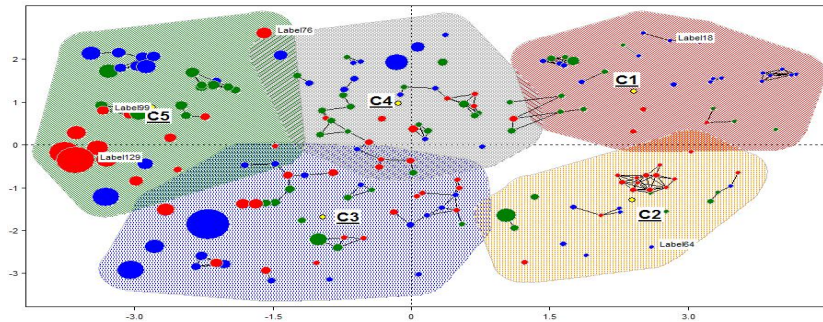


FIG. 7 – Visualization for each investment fund of its region - Europe (Blue), U.S. (green), World (Red) combined with its volatility (circles more or less large). The projection of a proximity network allows highlighting subclusters of very similar funds.

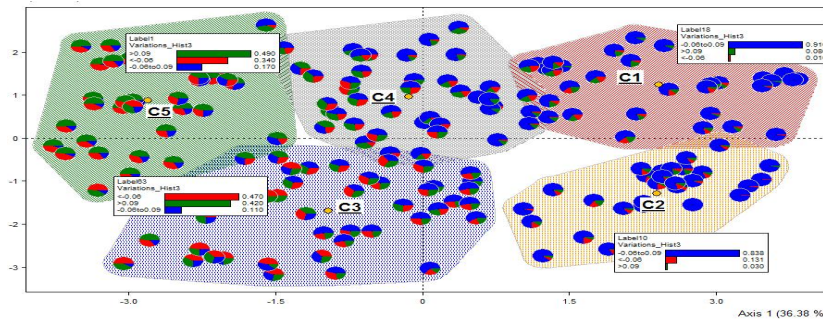


FIG. 8 – Projection, for each fund, of its daily variation "Variations\_Hist3" using pie charts. Clicking on a pie chart displays the histogram in detail. Daily variations amplify from the right to the left of the graph.

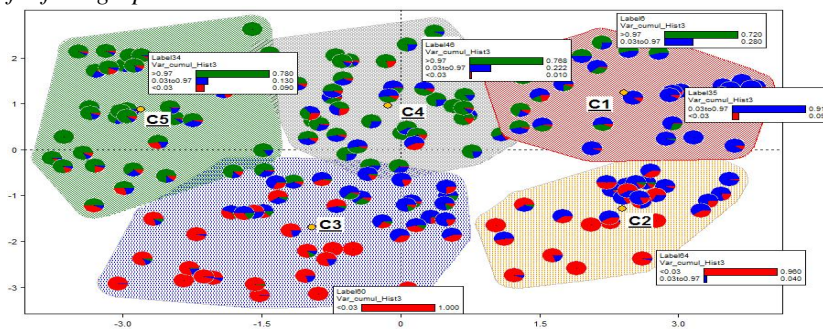


FIG. 9 – Projection, for each fund, of its daily cumulative returns "Var\_cumul\_Hist3" using pie charts. Clicking on a pie chart displays the histogram in detail. There is a clear separation between the clusters of profitable funds at the top of the graph and the unprofitable ones at the bottom.

In Figure 8, the values of the histogram variable "Variation\_Hist3" are displayed. These values are displayed as pie charts projected on the corresponding funds. Double-clicking on each pie chart, the details of the histogram will be highlighted. We find that the funds and clusters of funds have very low volatility on the right of the graph and become increasingly volatile as one moves to the left of the graph. Thus, the funds of clusters C1 and C2 are strongly characterized by steady daily returns (-0.06 to 0.09, the blue category). At the opposite, the funds of clusters C3 and C5, far left, are characterized by significant variations. These variations are both positive (green category > 0.09) and negative (red category < -0.06). Finally, we display the values of the histogram variable "Var\_cumul\_Hist3" on cumulative variations. A correlation between this variable and the second axis is observed. There is a clear separation between the clusters C2, C3 at the bottom of the graph with cumulative losses (or at the best very small gains) and the clusters C1, C4, C5 at the top with interesting cumulative gains. Thus, the cluster C1 is characterized by low volatility and slightly positive returns. C5 is a cluster of funds with high volatility but with very positive returns. C4 is an intermediate cluster. C2 contains funds with low volatility, but with a downward trend. Finally, cluster C3 is the riskiest one with volatile and unprofitable funds.

## 5 Conclusion

We have shown that SDA is a new paradigm based on the transition from standard individual observations to higher level observations described by symbolic data. SDA is a useful tool for Complex Data Mining. Much remains to be done for extending existing methods of Data Mining to Symbolic Data. A huge field of research is open as for example: the topology inside the symbolic space, summarizing graphs, social networks, extending to symbolic data with stochastic convergence Decision Trees, Factor Analysis, Canonical Analysis, PLS, Galois lattices, etc.. In the case of Big Data obtaining the symbolic data and analyzing them with scalable methods in the cloud is also a very open and interesting field of future research.

## References

- Afonso, F., E. Diday, N. Badez, Y. Genest, and A. Orcesi (2010). Use of symbolic data analysis for structural health monitoring applications. *In 2nd International Symposium on Life-Cycle Civil Engineering. IALCCE2010, Proc.Symp.Taipei, Taiwan. 27-31 October 2010.*
- Billard, L. and E. Diday (2006). *Symbolic Data Analysis: conceptual statistics and data Mining*. Chichester (England) Hoboken (NJ): Wiley Interscience. 321 pages.
- Blake, C. L. and C. J. Merz (1998). UCI Repository of machine learning databases. Technical report, University of California, Irvine, Dept. of Information and Computer Sciences.
- Diday, E. (1987). The symbolic approach in clustering and related methods of data analysis : the basic choices. *In 'Classification and Related Methods of Data Analysis', Proc. of IFCS'87, H.-H. Bock (ed.), Aachen, July 1987, North Holland, Amsterdam, 673-684.*
- Diday, E. (2010). Principal component analysis for categorical histogram data: Some open directions of research. *"Classification and Multivariate Analysis for Complex Data Structures" . B. Fichet, D. Piccolo, R. Verde, M. Vichi eds. 492 pages..*

## The symbolic data analysis paradigm

- Diday, E. (2011). Modélisation de données symboliques et application au cas des intervalles. *In proceeding of the SFC'2011. Orléans*, 119–122.
- Diday, E. and R. Emilion (2003). Maximal and stochastic Galois Lattices. *Journal of Discrete Applied Mathematics* 127. Elsevier (127), 271–284.
- Diday, E. and M. Noirhomme (2008). *Symbolic Data Analysis and the SODAS software*. Chichester (England) Hoboken (NJ): Wiley. 457 pages.
- Dougherty, J., R. Kohavi, and M. Sahami (1995). Supervised and unsupervised discretization of continuous features. *In Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufman Publishers, San Francisco, CA.*, 194–202.
- Douzal-Chouakria, A., L. Billard, and E. Diday (2011). Principal component analysis for interval-valued observations. *SAM (Statistical Analysis and Data Mining)* 4, 229–246.
- Fayyad, U. and K. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *In the 12th International Joint Conference on Artificial Intelligence, Morgan Kaufman Publishers, San Francisco, CA.*, 1022–1027.
- Fisher, W. (1958). On grouping for maximum of homogeneity. *JASA, Journal of the American Statistical Association*. 53, 789–798.
- Jolliffe, I. (2002). *Principal Component Analysis. Second Edition*. Springer Series in Statistics, New York. 487 pages.
- Kerber, R. (1992). Chimerge: Discretization for numeric attributes. *In proceedings of the ninth National Conference on Artificial Intelligence, Anaheim, California. AAAI Press / The MIT Press*, 123–128.
- Kuhn, T. (1962). The structure of scientific revolutions. *University of Chicago Press*.
- Le-Rademacher, J. and L. Billard (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference, volume 141, Issue 4*, 1593–1602.
- Quantin, C., L. Billard, M. Touati, N. Andreu, Y. Cottin, M. Zeller, F. Afonso, G. Battaglia, D. Seck, G. Le Teuff, and E. Diday (2011). Classification and regression trees on aggregate data modeling: An application in acute myocardial infarction. *Hindawi Publishing Corporation, Journal of Probability and Statistics* 2011.
- Ware, M. (2000). Weka documentation. Technical report, University of Weikoto.
- Yang, Y. and G. Webb (2001). Proportionnal k-interval discretization for naive-Bayes classifiers. *the 12th European Conference on Machine Learning*, 564 – 575.