The role of the Boxplot based Discretization in the conceptual interpretation of a hierarchical cluster

Karina Gibert, Alejandra Pérez-Bonilla
Departamento de Estadística e Investigación Operativa
Universitat Politécnica de Catalunya
Campus Nord, Edif. C5, C - Jordi Girona 1-3. 08034-Barcelona
karina.gibert@upc.edu, alejandra.perez@upc.edu

Abstract

In this paper special details that impact on good performance of *Boxplot based discretization (BbD)* are presented. The impact of improving *BbD* on the methodology *Conceptual characterization by embedded conditioning (CCEC)* of oriented to the automatic generation of conceptual descriptions of classifications that can support later decision-making is presented. Its application to the interpretation of the previously identified classes on a WasteWater Treatment Plant (WWTP). The particularity of the method is that it provides a conceptual interpretation of a partition in real domains or complex structure, on the basis of a previous hierarchical classification. The methodology is based on the use of some statistical tools (as the *multiple boxplot*, introduced by Tukey) to discover the structure of the data and to extract useful information.

1 Introduction

In a process of automatic clustering where the classes composing a certain domain are discovered, one of the most important required processes and one of the less standardized ones is, probably, that of *interpretation* of classes, closely related with the *validation* of classes, and critical in the later usefulness of the discovered knowledge [10]. The interpretation of the classes, so important to understand the meaning of the obtained clustering as well as the structure of the domain, use to be done in an artistic-like way. But this process becomes more and more complicate as the number of classes and involved variables grows. This work tries to face the problem of the automatic generation of interpretations of a classification, to develop, in the long term, a software supporting this task and to contribute to a more systematic interpretation process.

Boxplot based induction rules (BbIR), see [8], is a proposal to produce compact concepts associated to the classes, oriented to express the differential characteristic of every class in such a way that the user can easily understand which is the underlying classification criterion and can easily decide the treatment or action to be assigned to each class. Given a classification, the idea is to provide an automatic interpretation for it that supports the construction of intelligent decision support systems. The core of this process is a method for discretizing numerical variables in such a way that particularities of the classes are elicited called *Boxplot based discretization (BbD)*. In this work special details that impact on good performance of *BbIR* are presented.

A particular application to Waste Water Treatment Plants (WWTP) is in progress and results appear to be very promising. Examples used in this paper come for this real application. The presented proposal integrates different findings from a series of previous works, see [4], [7], in a single methodological tool which takes advantage of the hierarchical structure of the classification to overcome some of the limitations observed in [4], [12].

This is different from what is pursued by other

inductive learning techniques as association rules algorithms, see [2], where the set of produced association rules use to be huge in Data Mining context and the greater is the number of variables or/and classes in involved in the analysis, the more complex are the generated rules and the more difficult the interpretation of classes from those rules.

This paper is organized as follows: After the introduction, previous work of this research is in section §2. The section §3 presents Revising Boxplot based discretization. Finally in section §7 conclusions and future work.

2 Previous work

The present research is based on previous works in which the automatic process of characterization of classes has been analyzed. The main idea was to automatically analyze conditional distributions through *Multiple boxplot*, (see Figure 1) in order to identify *characterizing variables*, introduced in [4]:

> Let us consider $\mathcal{I}$ as the set of n objects to be analyzed. They are described by K numerical variables $X_k, (k = 1 : K); x_{ik}$ is the value taken by variable X_k for object i.

- Given a variable X_k and a partition $\mathcal{P}$ of $\mathcal{I}$, x is a *characterizing value* of class $C \in \mathcal{P}$ if $\exists i \in C$ tq $x_{ik} = x$ and $\forall i \notin C, x_{ik} \neq x$.
- Variable X_k is *Totally characterizing* class $C \in \mathcal{P}$, if either one or more of the values taken by X_k in class C are *characterizing values* of C.

Characterizing variables identify particularities of a given class, behaviors that are not observed out of that clases and therefore help to understand the *meaning* of a clustering.

The *Boxplot based discretization (BbD)* is presented in [7] as an efficient way of transforming a numerical variable into a qualitative one in such a way that the cut points for discretizing identify where the set of classes with non-null intersection of X_k changes and it consists on:

1. Calculate de *minimum* (m_C^k) and *maximum* (M_C^k) of X_k inside any class. Built $\mathcal{M}^k = \{m_{C_1}^k, \ldots, m_{C_\xi}^k, M_{C_1}^k, \ldots, M_{C_\xi}^k\}$, where $card(\mathcal{M}^k) = 2\xi$
2. Built the *set of cutpoints* $\mathcal{Z}^k$ by sorting $\mathcal{M}^k$ in increasing way into $\mathcal{Z}^k = \{z_i^k \ ; \ i = 1, \ldots, 2\xi\}$. At every z_i^k the set of class overlapping changes. In Fig.1, for example, both C_{391} and C_{389} take values between 30,592.1 and 52,255.8 but only C_{391}, takes values between 29,920.0 and 30,592.1 while only C_{389} between 30,592.1 and 54,088.6 .
3. Built the *set of intervals* I^k *induced by* $\mathcal{P}$ *on* X_k by defining an interval I_s^k between every pair of consecutive values of $\mathcal{Z}^k$. $I^k = \{I_1^k, \ldots, I_{2\xi-1}^k\}$ is the *BbD* of X_k. The I_s^k intervals have variable length and the set of intersecting classes is constant all along the interval and changes from one to another.

In [12] there is a preliminary proposal of building all the I_s^k following a single pattern: $I_s^k = (z_s^k, z_{s+1}^k]$ $\forall s > 1$ being $I_1^k = [z_1^k, z_2^k]$.

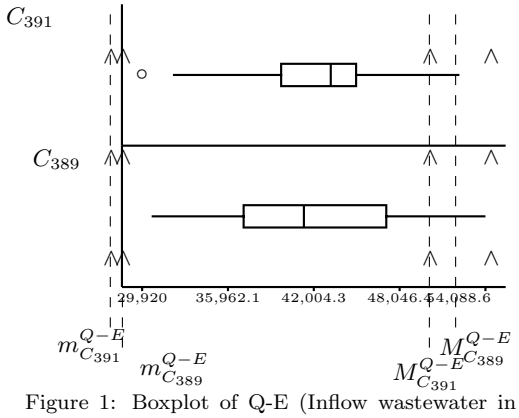

Figure 1: Boxplot of Q-E (Inflow wastewater in daily m^3 of water) in WWTP *vs* $\mathcal{P}_3$

The boxplot it is a graphical tool introduced by [11]. For each class the range of the variable is visualized and rare observations (outliers) are marked as "o" or "*". A box is displayed from Q1 (first quartile) to Q3 (third quartile) and the Median, usually inside the box, is marked with a vertical sign. Boxes include, then, the 50% of the elements of the class and the whiskers extend until the minimum and maximum.

In [6] a deeper discussion about situations in which closed or open intervals are more convenient is presented.

In [5] the formulation of the methodology *boxplot based induction rules (BbiR)* is presented. It is a method for generating probabilistic concepts with a minimum number of attributes on the basis of the *boxplot based discretization (BbD)* of X_k.

1. Use the *Boxplot based discretization* to build $\mathcal{I}^k = \{I_1^k, I_2^k, I_3^k, \ldots, I_{2\xi-1}^k\}$.
2. For every interval produce the rules: r_s : *If* $x_{ik} \in I_s^k \xrightarrow{p_{sC}} i \in C$
   where, $p_{sC} = P(C|I^k = I_s^k) = P(i \in C|x_{ik} \in I_s^k) = \frac{card\{i : x_{ik} \in I_s^k \wedge i \in C\}}{card\{i \in \mathcal{I} : x_{ik} \in I_s^k\}}$
   If $p_{sC} = 1$, I_s^k is a set of *characterizing values* of X_k. If $\forall s = 1 : 2\xi - 1$, $\exists C \in \mathcal{P}$ tq $p_{sC} = 1$ then X_k is a *totally characterizing variable*.

Although obtained results were satisfactory from an applied point of view, it is not clear that they are optimal in terms of coverage. This proposal was improved in [6] in such a way that the probability of the generated concepts increases, and yields more certain interpretations. The final goal of this research is to elaborate a new proposal, on the basis of previous work, which overcomes all the observed limitations and consolidates a methodology of automatic generation of interpretations from a given classification, giving support to the construction of intelligent decision support systems in the case of WWTP.

In [4] a comparison between a very primary version of the method and other inductive methods has shown that *BbIR* appears as a good *imitation* of the real process used by experts to manually interpret the classes. It also confirmed that modifying the method to provide more flexibility would sensibly improve its performance [5].

In [8] the *Boxplot based discretization (BbD)* was used in the *Methodology of conceptual characterization by embedded conditioning (CCEC)* which is a methodology for generating automatic interpretations of a given partition $\mathcal{P} \in \tau$, where $\tau = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, ..., \mathcal{P}_n\}$ is an indexed hierarchy of $\mathcal{I}$. Taking advantage of the hierarchy it is reduced to iteratively distinguish pairs of classes, what justifies that in this work only binary partitions are considered from now on.

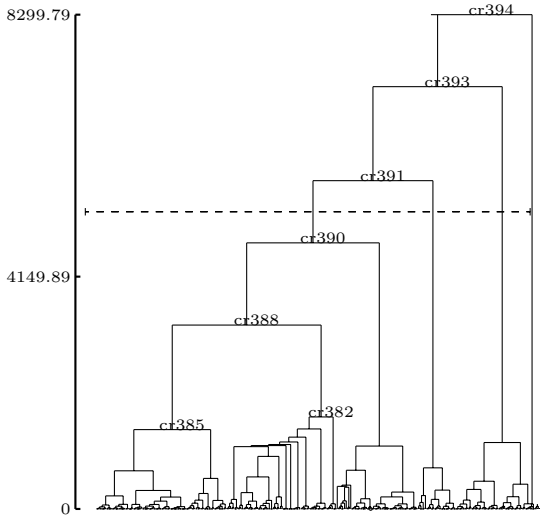

Figure 2: Hierarchical tree. $[\tau_{Gi2,R1}^{En,G}]$

2.1 Conceptual characterization by embedded conditioning (CCEC)

The Methodology *CCEC* [8],[9] which is a methodology for generating automatic interpretations of a given partition $\mathcal{P} \in \tau$, where $\tau = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, ..., \mathcal{P}_n\}$ is an indexed hierarchy of $\mathcal{I}$ which usually can be represented by a hierarchical binary tree, is present version it can only applied to a hierarchical clustering, since it uses the property of any binary hierarchical structure that $\mathcal{P}_{\xi+1}$ has the same classes of $\mathcal{P}_\xi$ except one, which splits in two subclasses in $\mathcal{P}_{\xi+1}$. This binary hierarchical structure is be used by *CCEC* to discover particularities of the final classes by an iterative process that search the variables distinguishing a pair of classes at a time, starting on the top of the hierarchical tree and going once step down at every iteration. Main idea is to analyze at every iteration the single pair of classes that split from previous iteration, the knowledge already obtained from the pathes of this pair of classes. Step by step also in hierarchical way. The *CCEC* [9] allows generation of automatic interpretations of a given partition $\mathcal{P}$ coming for a hierarchical tree. In §4 illustration of this particular process are introduced.

3 Revising Boxplot based discretization

For a binary partition $\mathcal{P}_2 = \{C_1, C_2\}$, Z^k always contains 4 elements which are minimum and maximum values of $C_1$and C_2 conveniently ordered. That is the reason why I^k will always have 3 intervals built upon Z^k values. In the particular case of binary partitions, the previous proposal [12], established the following structure for I^k: $I_1^k = [z_1^k, z_2^k], I_2^k = (z_2^k, z_3^k], I_3^k = (z_3^k, z_4^k]$

In [6] is evidenced that the rules generated from I^k following *BbIR* are sensitive to the form of the limits of every I_s^k.

classes	$min\ (m_C^{Q-E})$	$max\ (M_C^{Q-E})$
C_{391}^3	29,920.0	52,255.8
C_{389}^3	30,592.1	54,088.6

Table 1: Minimum and Maximum of $Q - E$.

$\mathcal{M}^{Q-E}$	$\mathcal{Z}^{Q-E}$
29,920.0	29,920.0
52,255.8	30,592.1
30,592.1	52,255.8
54,088.6	54,088.6

Table 2: $\mathcal{M}^{Q-E}$ and $\mathcal{Z}^{Q-E}$.

Let us analyse the simple example in figure 1, where the multiple boxplot of a variable Q-E vs a binary partition called $\mathcal{P}$ is displayed. Table 1 shows minimum and maximum of variable Q-E in classes C_{391} and C_{389}, while Table 2 shows $\mathcal{M}^{Q-E}$, set of extreme values of Q-E$|\mathcal{P}$ and $\mathcal{Z}^{Q-E}$, the corresponding sorting in increasing order. Following the previous proposal of [12], the $I^{Q-E} = \{I_1^{Q-E}, I_2^{Q-E}, I_3^{Q-E}\}$ is build in the following way: $I_1^{Q-E} = [29920.0, 30592.2]$, $I_2^{Q-E} = (30592.2, 52255.8]$, $I_3^{Q-E} = (52255.8, 54088.6]$.

From this, the *BbIR* produces the following set of rules:

$r_1: x_{i,Q-E} \in [29920.0, 30592.2] \xrightarrow{0.5} i \in C_{389}$

$r_2: x_{i,Q-E} \in [29920.0, 30592.2] \xrightarrow{0.5} i \in C_{391}$

$r_3: x_{i,Q-E} \in (30592.2, 52255.8] \xrightarrow{0.83} i \in C_{391}$

$r_4: x_{i,Q-E} \in (30592.2, 52255.8] \xrightarrow{0.17} i \in C_{389}$

$r_5: x_{i,Q-E} \in (52255.8, 54088.6] \xrightarrow{1.0} i \in C_{389}$

From the Figure 1 it is clear that z_s^k, and in consequence M_C^k, are identifying the points where class overlapping changes by a simple and cheap sorting of extreme values of conditioned distributions, which is extremely cheap compared with directly analysing intersections among classes with continuous variables; the area delimited by I_1^{Q-E} should certainly be assigned to C_{391} what is not according to r_1 and r_2. The reason why the probability of r_1 is 0.5 instead of 1 is that right limit of I_1^{Q-E} should be open instead of closed. Doing this redefinition, a new I^{Q-E} is defined as: $I_1^{Q-E} = [29920.0, 30592.2)$, $I_2^{Q-E} = [30592.2, 52255.8]$, $I_3^{Q-E} = (52255.8, 54088.6]$ and the new set of rules is:

$r_1: x_{i,Q-E} \in [29920.0, 30592.2] \xrightarrow{1.0} i \in C_{391}$

$r_2: x_{i,Q-E} \in (30592.2, 52255.8] \xrightarrow{0.827} i \in C_{391}$

$r_3: x_{i,Q-E} \in (30592.2, 52255.8] \xrightarrow{0.173} i \in C_{389}$

$r_4: x_{i,Q-E} \in (52255.8, 54088.6] \xrightarrow{1.0} i \in C_{389}$

Making a similar analysis on all different situations that can be found, Fig.3 shows the more convenient way of redefining the *Boxplot based discretization (BbD)* in each case, see [6] for details. Observing the column with header *pattern-CCEC* in Fig.3, it is easily seen that there are only 2 patterns of building I^k from Z^k according to the limits of this intervals. Both will generate I^k with 3 intervals:

1. **Open-*Center***: In the case 1 and case 2, 3 intervals are defined in such a way that the center (I_2^k) is an *open* interval by both sides. The pattern is the following: $I_1^k = [z_1^k, z_2^k]$, $I_2^k = (z_2^k, z_3^k)$ and $I_3^k = [z_3^k, z_4^k]$.
2. **Closed-*Center***: In the other cases (4 to 13), 3 intervals are defined in such a way that the center $[I_2^k]$ is a *closed* interval by both sides. The pattern is the following:$I_1^k = [z_1^k, z_2^k)$, $I_2^k = [z_2^k, z_3^k]$ and $I_3^k = (z_3^k, z_4^k]$.

These would represent a more realistic model for all different situations that can be found in the multiple Boxplot for two classes, detailed in [6] . None of these patterns coincides with the proposal in [12]. From this analysis it was also seen that the condition to generate an *open-center* pattern is: ($M_{C2}^k < m_{C1}^k$ and $M_{C1}^k < m_{C2}^k$), all other cases should be treated as *closed-center*. Using this new way of intervals generation it was seen that more certain rules can be induced from the classes which directly leads on more reliable interpretations. Table 4 shows the comparison between both proposals for some variables taken from the

previously referred real application on Waste-Water Treatment Plants (WWTP), which is in progress. In most of the cases the number of rules with probability 1 produced by *Revised BbD* increases (rules with null probability can be eliminated).

Case	Characteristic	Multiple Boxplot	r*BbD*
1	$M^k_{C2} < m^k_{C1}$		[ ],( ), [ ]
2	$M^k_{C1} < m^k_{C2}$		[ ],( ), [ ]
3	$m^k_{C1} = m^k_{C2} \wedge M^k_{C1} = M^k_{C2}$		[ ), [ ], ( ]
4	$m^k_{C1} > m^k_{C2} \wedge M^k_{C1} > M^k_{C2} \wedge m^k_{C1} < M^k_{C2}$		[ ), [ ], ( ]
5	$m^k_{C1} < m^k_{C2} \wedge M^k_{C1} < M^k_{C2} \wedge m^k_{C2} > M^k_{C1}$		[ ), [ ], ( ]
6	$m^k_{C1} < m^k_{C2} \wedge M^k_{C1} > M^k_{C2} \wedge m^k_{C1} < M^k_{C2}$		[ ), [ ], ( ]
7	$m^k_{C1} > m^k_{C2} \wedge M^k_{C1} < M^k_{C2} \wedge m^k_{C1} < M^k_{C2}$		[ ), [ ], ( ]
8	$m^k_{C1} = m^k_{C2} \wedge M^k_{C1} < M^k_{C2}$		[ ), [ ], ( ]
9	$m^k_{C1} = m^k_{C2} \wedge M^k_{C1} > M^k_{C2}$		[ ), [ ], ( ]
10	$m^k_{C1} > m^k_{C2} \wedge M^k_{C1} = M^k_{C2}$		[ ), [ ], ( ]
11	$m^k_{C1} < m^k_{C2} \wedge M^k_{C1} = M^k_{C2}$		[ ), [ ], ( ]
12	$M^k_{C1} = m^k_{C2}$		[ ), [ ], ( ]
13	$M^k_{C2} = m^k_{C1}$		[ ), [ ], ( ]

Figure 3: Análisis Descriptivo por clases para $\mathcal{P}_2$

4 Application

In this section the use of *CCEC* to generate the domain decision model from the final cluster is shown.

4.1 The target domain

The main goal of wastewater treatment plants is, in general terms, to guarantee the outflow water quality (referred to certain legal requirements), in order to restore the natural environmental balance which is disturbed by industry wastes, domestic waste-waters, etc. When the plant is not on normal operation, which is extremely difficult to model by traditional mechanicistic models, decisions have to be taken to modify some parameters of the wastewater treatment process in order to reestablish the normality as soon as possible. This process is very complex, because of the intrinsic features of wastewater, because of the bad consequences of an incorrect management of the plant. A very brief description of the process in the plant is presented: the water flows sequentially through three or four processes which are commonly known as pretreatment, primary, secondary, and advanced treatment (see [1] for a detailed description of the process). Figure 4 (right) depicts the general structure of the plant.

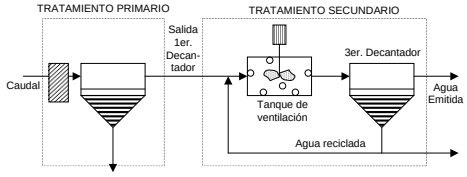

Figure 4: General structure of the WWTP

4.2 The database

Data analyzed in this paper comes from a catalan WWTP, in Spain. It is a sample of 396 observations taken from September the first of 1995 to September the 30th of 1996. Each observation refers to a daily mean, and it is identified by the date itself. The state of the Plant is described through a set of 25 variables, considered the more relevant upon expert's opinions. They can be grouped[1] as Table 3.

4.3 Reference clustering

The data base presented in §4.2, was classified in a previous work [7] by using automatic methods, producing the hierarchical tree of Figure 2. As usual in hierarchical clustering, the final partition is the horizontal cut of the tree that maximizes the ratio between *heterogeneity* between classes with respect to *homogeneity* within classes, what guaranties the *distinguishability* between classes. The result is a 4 classes partition $\mathcal{P}_4 = \{C_{392}, C_{389}, C_{390}, C_{383}\}$.

[1] Measurement Points: Input(E), After Settler(D), Biologic Treatment(B), Output(S). Other variables: Recirculated Flow(QRG), Purged flow(QPG) and Air inflow(QAG).

VARIABLES	E	D	B	S
Inflow(Q)	Q-E		QB-B	
Iron Pre-treatment(FE)	FE-E			
Hydrogen Potential(PH)	PH-E	PH-D		PH-S
Suspended Solids(SS)	SS-E	SS-D		SS-S
Volatile Suspended Solids(SSV)	SSV-E	SSV-D		SSV-S
Chemical Organic Matter(DQO)	DQO-E	DQO-D		DQO-S
Biodegradable Organic Matter(DBO)	DBO-E	DBO-D		DBO-S
Index 30 at the Biological Reactor(V30)			V30-B	
Mixed Liquor Suspended Solids(MLSS)			MLSS-B	
Mixed Liquor Volatile Suspended Solids(MLVSS)			MLVSS-B	
Mean Cell Residence Time(MCRT)			MCRT-B	

Table 3: Variables used in the Clustering

It.	Original Boxplot based discretization	Revised Boxplot based discretization
1	$r_1: x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} C_{392}$	$r_1: x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} C_{392}$
	$r_2: x_{Q-E,i} \in (23662.9, 29920.0] \xrightarrow{1.0} C_{393}$	
	$r_3: x_{Q-E,i} \in (29920.0, 54088.6] \xrightarrow{1.0} C_{393}$	$r_3: x_{Q-E,i} \in [29920.0, 54088.6] \xrightarrow{1.0} C_{393}$
2	$r_1: x_{Q-E,i} \in [29920.0, 30592.2] \xrightarrow{0.5} C_{389}$	
	$r_2: x_{Q-E,i} \in [29920.0, 30592.2] \xrightarrow{0.5} C_{391}$	$r_2: x_{Q-E,i} \in [29920.0, 30592.2) \xrightarrow{1.0} C_{391}$
	$r_3: x_{Q-E,i} \in (30592.2, 52255.8] \xrightarrow{0.17} C_{389}$	$r_3: x_{Q-E,i} \in [30592.2, 52255.8] \xrightarrow{0.17} C_{389}$
	$r_4: x_{Q-E,i} \in (30592.2, 52255.8] \xrightarrow{0.83} C_{391}$	$r_4: x_{Q-E,i} \in [30592.2, 52255.8] \xrightarrow{0.83} C_{391}$
	$r_5: x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} C_{389}$	$r_5: x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} C_{389}$
	$r_1: x_{QR-G,i} \in [26218.0, 27351.0] \xrightarrow{0.67} C_{389}$	$r_1: x_{QR-G,i} \in [26218.0, 27351.0) \xrightarrow{1.0} C_{389}$
	$r_2: x_{QR-G,i} \in [26218.0, 27351.0] \xrightarrow{0.33} C_{391}$	
	$r_3: x_{QR-G,i} \in (27351.0, 43298.1] \xrightarrow{0.21} C_{389}$	$r_3: x_{QR-G,i} \in [27351.0, 43298.1] \xrightarrow{0.21} C_{389}$
	$r_4: x_{QR-G,i} \in (27351.0, 43298.1] \xrightarrow{0.79} C_{391}$	$r_3: x_{QR-G,i} \in [27351.0, 43298.1] \xrightarrow{0.79} C_{391}$
	$r_5: x_{QR-G,i} \in (43298.1, 49527.0] \xrightarrow{1.0} C_{391}$	$r_5: x_{QR-G,i} \in (43298.1, 49527.0] \xrightarrow{1.0} C_{391}$
3	$r_1: x_{DBO-D,i} \in [36.0, 56.0] \xrightarrow{0.88} C_{389}$	$r_1: x_{DBO-D,i} \in [36.0, 56.0) \xrightarrow{1.0} C_{389}$
	$r_2: x_{DBO-D,i} \in [36.0, 56.0] \xrightarrow{0.13} C_{391}$	
	$r_3: x_{DBO-D,i} \in (56.0, 171.0] \xrightarrow{0.2} C_{389}$	$r_3: x_{DBO-D,i} \in [56.0, 171.0] \xrightarrow{0.2} C_{389}$
	$r_4: x_{DBO-D,i} \in (56.0, 171.0] \xrightarrow{0.8} C_{391}$	$r_4: x_{DBO-D,i} \in [56.0, 171.0] \xrightarrow{0.8} C_{391}$
	$r_5: x_{DBO-D,i} \in (171.0, 274.0] \xrightarrow{1.0} C_{391}$	$r_5: x_{DBO-D,i} \in (171.0, 274.0] \xrightarrow{1.0} C_{391}$
	$r_1: x_{DBO-E,i} \in [90.00, 220.0] \xrightarrow{0.96} C_{390}$	$r_1: x_{DBO-E,i} \in [90.00, 220.0) \xrightarrow{1.0} C_{390}$
	$r_2: x_{DBO-E,i} \in [90.00, 220.0] \xrightarrow{0.01} C_{383}$	
	$r_3: x_{DBO-E,i} \in (220.0, 382.0] \xrightarrow{0.99} C_{390}$	$r_3: x_{DBO-E,i} \in [220.0, 382.0] \xrightarrow{0.78} C_{390}$
	$r_4: x_{DBO-E,i} \in (220.0, 382.0] \xrightarrow{0.22} C_{383}$	$r_4: x_{DBO-E,i} \in [220.0, 382.0] \xrightarrow{0.22} C_{383}$
	$r_5: x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} C_{383}$	$r_5: x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} C_{383}$

Table 4: Comparison between both proposal.

5 Impact of the *Revised BbD*

In [8] the original version of the *BbD* was applied to obtain one automatic interpretation of $\mathcal{P}_4$ using. In this work, *CCCS* with the revised *BbD* has been applied and improvements on the final interpretation are shown in the following.

For the first iteration in that C_{392}*separates* of $C_{393} = C_{390} \cup C_{389} \cup C_{383}$ the totally characterizing variables continues being the same ones in 2 versions of the *BbD* and with those we have stayed. However all the systems of rules of the totally characterizing variables wins pre-

cision with the revised version. We present the case of Q-E to way illustration; this is a totally characterizing variable, to see Figure 5(left) and you it uses to separate C_{392} of C_{393}, but with the *original BbD* the interval I_2^{Q-E}(to see Table 4 left column It.1) it generated a sure rule that contained in reality an only point (29.920) and in fact it assigned with probability 1 to C_{393} the whole tract of 23.662,9 at 29.920 when in fact that is an empty tract. In the revised version of *BbD* it is corrected this anomaly and the point 29920 you it incorporates r_3 being I_2^{Q-E} empty and therefore unnecessary. This produces improvements in the interpretation of those variables that are to high values that it is the case of Q-E, QB-B, QP-G, QR-G and MCRT-B for C_{392}.

In the second iteration, we separate $C_{391} = C_{390} \cup C_{383}$ of C_{389} to be this a subdivision of C_{393}, we know that, as characteristics shared with C_{391}, won't have the flows of entrance, of recirculation and purge low, the cellular age not it will be high, what distinguishes them of C_{392}. Of the 5 variables totally characterizing of C_{392} 2 of them (Q-E and QB-B) C_{389} doesn't discriminate against of C_{391} with the *original BbD* and the other ones 3 made it with uncertainty. In all these cases the revision of the method produces certain rules ($p = 1$) that the one increases discrimination power. Let us see what happened with Q-E (and QB-B) for example. The Table left 4(column It.2) it shows the rules obtained with *original BbD*. Of the iteration 1 were already known that very low flows $[20500.0, 23662.9]m^3$ they went to C_{392}. With the rules of the It.2 izq. the only thing that one can say is that C_{389} has flows non low the same as C_{391}, but the *revised BbD* it allows clearly to discern that C_{389} has high flows $(52255.8, 54088.6]m^3$ and C_{391} the intermissions. The case of those other 3 variables illustrate it QR-G: we observe that r_1 it becomes certain and specific what we knew about the It.1 (C_{393} it was not low and here it is seen that takes bigger values that in C_{392} and smaller than in C_{391} in C_{389}).

With the *original BbD* for to interpret it is used; QB-B, MCRT-B and MLSSV-B and they only existed 3 certain rules that assigned to C_{389}. The 3 rules pointed at high levels and they continue being safe with the *revised BbD* but a good group of variables that they pointed also exists to value low and they have passed to generate certain rules with the *revised BbD*, as the case of the variable DBO-D that before had $p = 0.88$ and it passes at 1, that is to say of 1 certain rule and 4 uncertain we pass at 2 certain and 2 uncertain, to see Table 4(columna It.2); and it presents a similar behavior to the one of Q-E in the It.1 (I_1^{DBO-D} it only contains an erroneous point). In the same case of DBO-D you they find other 4 variables that $p > 0.75had$ with the *original BbD* what had allowed to consider them in the interpretation, but for other 4 variables rules appear certain that before had $0.5 < p < 0.75$ and they incorporate to the interpretation. Also, at the request of the experts 3 were used additional illustrative variables to enrich the interpretation of C_{389} that was a so much poor person that with the *revised BbD* they go to $p = 1$ again.

In the third and last iteration,we separate C_{390} of C_{383} and the general situation is the following: With the *original BbD* 5 variables only allowed a clear discrimination between both classes since all the other variables produce groups of rules with C_{390} as only part right. The *revised BbD* it produces an effect resemblance and the 5 variables that discriminate against are the same ones. However with the *revised BbD* the rules that go to C_{390} and those that go to C_{383} they are safer. By way of illustration we observe what passes with the variable DBO-E, to the being 220 the minimum of C_{383} and to have defined I_1^{DBO-E} closed to the right that point is shared among r_1 and r_3 increasing the uncertainty in the assignments, to see Table 4. When opening up I_1^{DBO-E} to the right makes that r_1 it becomes certain and r_2 it becomes impossible and the addition of a single point to r_1 hardly has impact about the certainty of those rules. This way we pass of 4 you not rule certain and 1 certain starting from 3 intervals not holes 2 certain rules and 2 not certain, with that which improves the interpretive capacity of the system, we have applied the pattern Closed *centro forms*.

6 Results

Class C_{392}: In fact this class has several totally characterizing variables, most of them at low values. After a global look of those variables, and checking the dates corresponding to days assigned to this class, experts identified this as a class of storming days. To face the storm, the decision of plant head was to minimize the water inflow (Q-E), closing the input valves, and to maintain the microorganisms (by minimizing the purged flow (QP-G)) in order to protect the system. This produce increasing of biomass in the reactor and justify the hight values for MCRT-B, see Table 5.

Class C_{389}: Whit the *original BbD* these days the plant works very well even reduces ammonium (NH4-D y NKT-D lows) in spite of not being a plant specifically designed for this. In addition, the reactor is working a full yield (QB-B high) and the degradation is good (MCRT-B and CM2-B low), as a result the water is so clean (DBO-S is low). Experts identify this class with a profile of excellent working of the plant. Whit the *revised BbD* analyzed in this work, we observed that the water he/she is entering quite clean (SSV-AND, DBO-AND low) that the settler also works well and that there is not that to purge neither to recycle (QA-G, QR-G, QP-G low) what corroborates the excellent operation of the biological reactor, see Table 5.

Class C_{383}: It is only to distinguish C_{383} of C_{390}knowing that both they inherit the characteristics of C_{393}. they Indicate that the water that enters it is very dirty (SS-AND, SSV-E, DQO-E and DBO-E high) so much in organic matter as in solids in suspension. According to the expert a load crash exists (organic matter) of solids in the entrance of the process, although the plant is working, as it indicates the fact that the volatile solids in suspension inside of the biologic reactor (MLVSS-B) they are low. The revision of the method it only increases the certainty of the last variable, see Table 5.

Class C_{390}: This class is characterized for to have the complementary characteristics of the previous ones, with that which, we are speaking of days in those that the flows of entrance is not very low, the entrance of the reactor is normal, the cellular age is not very high, the operation is not so good as to even eliminate the ammonium. In fact the assigned days to this class they present some punctual problem since, although the water that neither arrives is very dirty (like they indicate the values low or means of SS-E, SSV-E, DQO-E and DBO-E), the solids in suspension in the biological reactor stays more high of him that it would correspond (MLVSS-B), what indicates that the purification not it is good. 4 of the rules have increased their probability up to 1 and the other ones have been same. The list of more probable rules that they go to C_{390}, in Table 5.

7 Conclusions and future work.

In this paper a revision of *Boxplot based discretization (BbD)* is presented in such a way that the resulting discretization of a numerical variable allows induction of more certain rules. The *BbD* is a step of a wider methodology called *CCEC* [8] which is oriented to generate automatic interpretations from a group of classes in such a way that concepts associated to classes are built taking advantage of hierarchical structure of the underlying clustering.

The *BbD*, is a quick and effective method for discretizing numerical variables for generating conceptual model of the domain, which will greatly support the posterior decision-making. *Revised BbD* has been included in *CCEC* and the whole methodology has been successfully applied this to real data coming from a Wastewater Treatment Plant. Benefits of this proposal are of special interest in the interpretation of partitions with great number of classes. The induced model can be included as a part of an Intelligent Decision Support System to recommend decisions to be taken in a certain new situation. The main requirement of *CCEC* is to have an efficient way of discretizing numerical variables according to the subsets of classes that can share values of a certain variable, so the generated concepts can express particularities than can distinguish one class from the others. The main property of *Revised BbD* is that it allows finding the

Original Boxplot based discretization	Revised Boxplot based discretization
Class C_{392}	
$r^{Q-E}_{1,C_{392}} : x_{Q-E,i} \in [20500.0, 23662.900] \xrightarrow{1.0} C_{392}$	$r^{Q-E}_{1,C_{392}} : x_{Q-E,i} \in [20500.0, 23662.900] \xrightarrow{1.0} C_{392}$
$r^{QB-B}_{1,C_{392}} : x_{QB-B,i} \in [19883.0, 22891.00] \xrightarrow{1.0} C_{392}$	$r^{QB-B}_{1,C_{392}} : x_{QB-B,i} \in [19883.0, 22891.00] \xrightarrow{1.0} C_{392}$
$r^{MCRT-B}_{2,C_{392}} : x_{MCRT-B,i} \in (34.4, 179.8] \xrightarrow{1.0} C_{392}$	
$r^{MCRT-B}_{3,C_{392}} : x_{MCRT-B,i} \in (179.8, 342] \xrightarrow{1.0} C_{392}$	$r^{MCRT-B}_{3,C_{392}} : x_{MCRT-B,i} \in [179.8, 342] \xrightarrow{1.0} C_{392}$
$r^{QR-G}_{1,C_{392}} : x_{QR-G,i} \in [17932.6, 18343.50] \xrightarrow{1.0} C_{392}$	$r^{QR-G}_{1,C_{392}} : x_{QR-G,i} \in [17932.6, 18343.50] \xrightarrow{1.0} C_{392}$
$r^{QP-G}_{1,C_{392}} : x_{QP-G,i} \in [0.00000, 0.000000] \xrightarrow{1.0} C_{392}$	$r^{QP-G}_{1,C_{392}} : x_{QP-G,i} \in [0.00000, 0.000000] \xrightarrow{1.0} C_{392}$
Class C_{389}	
$r^{Q-E}_{3,C_{389}} : x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} C_{389}$	$r^{Q-E}_{3,C_{389}} : x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} C_{389}$
	$r^{SSV-E}_{1,C_{389}} : x_{SSV-E,i} \in [30.0, 60.0) \xrightarrow{1.0} C_{389}$
	$r^{DBO-E}_{1,C_{389}} : x_{DBO-E,i} \in [69.0, 90.0) \xrightarrow{1.0} C_{389}$
$r^{NKT-D}_{1,C_{389}} : x_{NKT-D,i} \in [22.9, 27.4] \xrightarrow{0.67} C_{389}$	$r^{NKT-D}_{1,C_{389}} : x_{NKT-D,i} \in [22.9, 27.4) \xrightarrow{1.0} C_{389}$
$r^{NH4-D}_{1,C_{389}} : x_{NH4-D,i} \in [13.3, 17.4] \xrightarrow{0.67} C_{389}$	$r^{NH4-D}_{1,C_{389}} : x_{NH4-D,i} \in [13.3, 17.4) \xrightarrow{1.0} C_{389}$
	$r^{DBO-D}_{1,C_{389}} : x_{DBO-D,i} \in [36.0, 56.0) \xrightarrow{1.0} C_{389}$
	$r^{SS-D}_{1,C_{389}} : x_{SS-D,i} \in [48.0, 63.0) \xrightarrow{1.0} C_{389}$
	$r^{SSV-D}_{1,C_{389}} : x_{SSV-D,i} \in [30.0, 47.0) \xrightarrow{1.0} C_{389}$
	$r^{PH-D}_{1,C_{389}} : x_{PH-D,i} \in [7.1, 7.2) \xrightarrow{1.0} C_{389}$
$r^{QB-B}_{3,C_{389}} : x_{QB-B,i} \in (49695.8, 52244.6] \xrightarrow{1.0} C_{389}$	$r^{QB-B}_{3,C_{389}} : x_{QB-B,i} \in (49695.8, 52244.6] \xrightarrow{1.0} C_{389}$
$r^{MLSS-B}_{3,C_{389}} : x_{MLSS-B,i} \in (2696, 2978] \xrightarrow{1.0} C_{389}$	$r^{MLSS-B}_{3,C_{389}} : x_{MLSS-B,i} \in (2696, 2978] \xrightarrow{1.0} C_{389}$
$r^{MCRT-B}_{3,C_{389}} : x_{MCRT-B,i} \in (28.8, 34.4] \xrightarrow{1.0} C_{389}$	$r^{MCRT-B}_{3,C_{389}} : x_{MCRT-B,i} \in (28.8, 34.4] \xrightarrow{1.0} C_{389}$
	$r^{CM2-B}_{1,C_{389}} : x_{CM2-B,i} \in [0.06, 0.1) \xrightarrow{1.0} C_{389}$
	$r^{DBO-S}_{1,C_{389}} : x_{DBO-S,i} \in [2.0, 4.0) \xrightarrow{1.0} C_{389}$
	$r^{QA-G}_{1,C_{389}} : x_{QA-G,i} \in [124120, 136371) \xrightarrow{1.0} C_{389}$
	$r^{QR-G}_{1,C_{389}} : x_{QR-G,i} \in [26218.0, 27351.0) \xrightarrow{1.0} C_{389}$
	$r^{QP-G}_{1,C_{389}} : x_{QP-G,i} \in [188.0, 327.6) \xrightarrow{1.0} C_{389}$
Class C_{383}	
$r^{DBO-E}_{3,C_{383}} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} C_{383}$	$r^{DBO-E}_{3,C_{383}} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} C_{383}$
$r^{SSV-E}_{3,C_{383}} : x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{1.0} C_{383}$	$r^{SSV-E}_{3,C_{383}} : x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{1.0} C_{383}$
$r^{DQO-E}_{3,C_{383}} : x_{DQO-E,i} \in (1279, 1579] \xrightarrow{1.0} C_{383}$	$r^{DQO-E}_{3,C_{383}} : x_{DQO-E,i} \in (1279, 1579] \xrightarrow{1.0} C_{383}$
$r^{SS-E}_{3,C_{383}} : x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{1.0} C_{383}$	$r^{SS-E}_{3,C_{383}} : x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{1.0} C_{383}$
$r^{MLVSS-B}_{1,C_{383}} : x_{MLVSS-B,i} \in [185, 611] \xrightarrow{0.67} C_{383}$	$r^{MLVSS-B}_{1,C_{383}} : x_{MLVSS-B,i} \in [185, 611) \xrightarrow{1.0} C_{383}$
Class C_{390}	
$r^{DBO-E}_{1,C_{390}} : x_{DBO-E,i} \in [90.00, 220.0] \xrightarrow{0.99} C_{390}$	$r^{DBO-E}_{1,C_{390}} : x_{DBO-E,i} \in [90.00, 220.0) \xrightarrow{\mathbf{1.0}} C_{390}$
$r^{SSV-E}_{1,C_{390}} : x_{SSV-E,i} \in [60.00, 92.0] \xrightarrow{0.88} C_{390}$	$r^{SSV-E}_{1,C_{390}} : x_{SSV-E,i} \in [60.00, 92.0) \xrightarrow{\mathbf{1.0}} C_{390}$
$r^{SSV-E}_{2,C_{390}} : x_{SSV-E,i} \in (92.0, 336.0] \xrightarrow{0.89} C_{390}$	$r^{SSV-E}_{2,C_{390}} : x_{SSV-E,i} \in [92.0, 336.0] \xrightarrow{0.89} C_{390}$
$r^{DQO-E}_{1,C_{390}} : x_{DQO-E,i} \in [158.0, 414] \xrightarrow{0.99} C_{390}$	$r^{DQO-E}_{1,C_{390}} : x_{DQO-E,i} \in [158.0, 414) \xrightarrow{\mathbf{1.0}} C_{390}$
$r^{DQO-E}_{2,C_{390}} : x_{DQO-E,i} \in (414, 1279] \xrightarrow{0.85} C_{390}$	$r^{DQO-E}_{2,C_{390}} : x_{DQO-E,i} \in [414, 1279] \xrightarrow{0.84} C_{390}$
$r^{SS-E}_{1,C_{390}} : x_{SS-E,i} \in [82.00, 114.0] \xrightarrow{0.8} C_{390}$	$r^{SS-E}_{1,C_{390}} : x_{SS-E,i} \in [82.00, 114.0) \xrightarrow{\mathbf{1.0}} C_{390}$
$r^{SS-E}_{2,C_{390}} : x_{SS-E,i} \in (114.0, 480.0] \xrightarrow{0.89} C_{390}$	$r^{SS-E}_{2,C_{390}} : x_{SS-E,i} \in [114.0, 480.0] \xrightarrow{0.89} C_{390}$
$r^{MLVSS-B}_{2,C_{390}} : x_{MLVSS-B,i} \in (611, 1726] \xrightarrow{0.89} C_{390}$	$r^{MLVSS-B}_{2,C_{390}} : x_{MLVSS-B,i} \in (611, 1726] \xrightarrow{0.89} C_{390}$
$r^{MLVSS-B}_{3,C_{390}} : x_{MLVSS-B,i} \in (1726, 2054] \xrightarrow{\mathbf{1.0}} C_{390}$	$r^{MLVSS-B}_{3,C_{390}} : x_{MLVSS-B,i} \in (1726, 2054] \xrightarrow{\mathbf{1.0}} C_{390}$

Table 5: Induction Rules for each interpreting class.

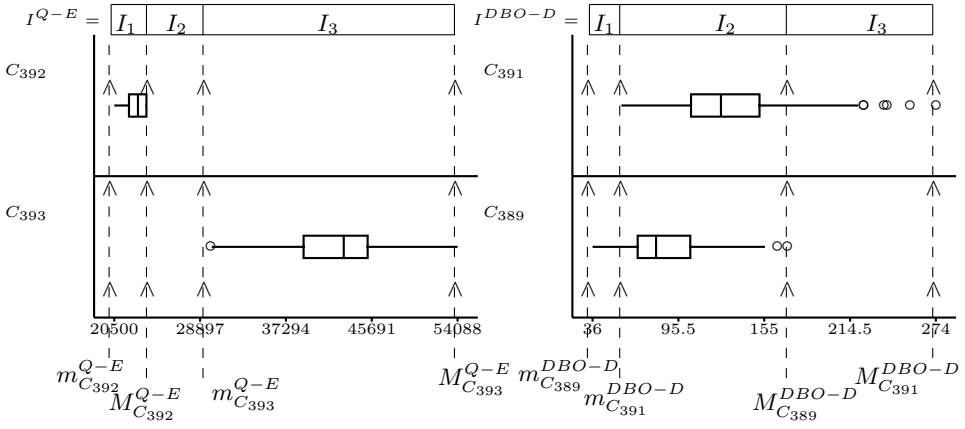

Figure 5: Multiple Boxplot : Q-E vs $\mathcal{P}_2$(left); DBO-D vs $\mathcal{P}_3 - \{C_{392}\}$ (right)

cut points of the variable where class overlapping changes by a simple and cheap sorting of extreme values of conditioned distributions, which is extremely cheap compared with directly analysing intersections among classes with continuous variables. The *Revised BbD* increases the quality of produced knowledge and decision making support incrementing the number of certain rules. At present, different criteria for deciding which variable is to be kept in the final interpretation are being analysed to see the impact of *BbD* on final interpretation. As Fayyad [3] , points out, Data Mining should also be involved with "interpretation of the patterns generated by Data Mining algorithms". *CCEC* (and *Revised BbD*) is trying to contribute to this particular issue.

References

[1] Abrams and Eddy. Wastewater engineering treatament, disposal,reuse. 4th Ed. revised by George Tchobanoglous, Franklin L. Burton NY.US. McGraw-Hill., 2003.

[2] Agraval, R. and Sikrant, R. Fast algorithms for mining association rules. In: VLDB94. Procs. OF 20th Irtl Conf. on Very large Data Bases. 487493. 1994.

[3] Fayyad, U. et alt. From Data Mining to Knowledge Discovery: An overview. Advances in KD and DM. AAAI/MIT Press. pp 1–34. 1996.

[4] Gibert, K. The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Communications*, 9(1):36–37, march 1996.

[5] Gibert, K. *Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos*. Red Nacional de MiDA, 2004.

[6] Gibert, K. and Pérez-Bonilla, A. Analsis y propiedades de la Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS). Research report DR 2005/14 prensa, UPC, 2005.

[7] Gibert, K. and Roda, I. Identifying characteristic situations in wastewater treatment plants. In *Workshop BESAI* (ECAI2000), v. 1, pp. 1–9. 2000.

[8] Gibert, K. and Pérez-Bonilla, A. Automatic generation of interpretation as a tool for modelling decisions. III International Conference on Modeling Decisions for Artificial Intelligence, 2006.

[9] Gibert, K. and Pérez-Bonilla, A. Ventajas de la estructura jerárquica del clustering en la interpretación automática de clasificaciones. III TAMIDA Workshop, Granada, THOMPSON 67-76. 2005.

[10] Gordon, A. D. Identifying genuine clusters in a classification. Computational. Statistics and Data Analysis. V.18: 561-581. 1994.

[11] Tukey, J.W. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[12] Vázquez, F. and Gibert,K. Generación Automática de Reglas Difusas en Dominios Poco Estructurados con Variables Numéricas. In Proc. CAEPIA V.I, pp 143–152. 2001.