Jul 1st, 12:00 AM

Automatic generation of conceptual descriptions of classifications in Environmental Domains

Alejandra Perez-Bonilla

Karina Gibert

Darko Vrecko

iEMSs 2008:International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Automatic generation of conceptual descriptions of classifications in Environmental Domains

Alejandra Perez-Bonilla [a], **Karina Gibert**[a] **and Darko Vrecko**[b]

[a]*Department of Statistics and Operations Research. Technical University of Catalonia, Edif. C5, Jordi Girona 1-3, 08034 Barcelona, Spain (alejandra.perez@upc.edu, karina.gibert@upc.edu)*

[b]*Department of Systems and Control. Jozef Stefan Institute, Jamova 39. Ljubljana. Slovenia (darko.vrecko@ijs.si)*

Abstract: In this paper the *Methodology of conceptual characterization by embedded conditioning CCEC*, oriented to the automatic generation of conceptual descriptions of classifications that can support later decision-making in Environmental Domains is applied to the interpretation of previously identified classes characterizing situations on a Waste Water Treatment Plant (WWTP). The particularity of the method is that it provides an interpretation of a partition previously obtained on an *ill-structured domain*, on the basis of a previous hierarchical clustering. The methodology uses some statistical tools (such as the *multiple boxplot*) together with artificial intelligent tools (as some machine learning methods), to learn the structure of the data; this allows extracting useful information (using the concept of *characterizing variable*) for the automatic generation of a set of useful concepts for later identification of classes. In this paper the usefulness of *CCEC* for building domain theories as models for supporting later decision-making is addressed and contrasted with interpretation provided by experts.

Keywords: Knowledge Discovery and Data Mining, Hierarchical clustering, class interpretation, Induction rules, Waste Water treatment plants.

1 INTRODUCTION

In automatic classification where the classes composing a certain domain are to be discovered, one of the most important required processes and one of the less standardized, is the interpretation of the classes (Gordon [1994]), closely related with *validation*, and critical in the later usefulness of the discovered knowledge. The interpretation of the classes, so important to understand the meaning of the obtained classification as well as the structure of the domain, used to be done in an artistic-like way. But this process becomes more and more complicated as the number of classes grows. This work is involved with the automatic generation of useful interpretations of classes in such a way that decisions about the action associated to a new object can be modelled and it is oriented to develop, in the long term, decision support system.

The presented proposal integrates different findings from a series of previous works; Gibert [1996], Pérez-Bonilla and Gibert [2007] proposed a single methodological tool which takes advantage of the hierarchical structure of the clustering to overcome some of the limitations observed in Gibert [1996], Gibert et al. [1998], Gibert and Pérez-Bonilla [2006]. This paper is organized as follows. After the introduction, basics concepts are presented in §2, on order to be understand the methodology which is presented in §3. WWTP as well as description of the specific data base that has been analyzed are presented in §4. Results of applying *CCEC* to the data described are given in §5. Finally in §6 the conclusions and the future work are addressed.

2 BASIC CONCEPTS

Four main concepts of methodology CCEC that are used in this work (basic notation is introduced in Figure 1(left)) are:

- Support (Sup): Given a rule $r : x_{ik} \in I_s^k \longrightarrow C$, where $I_s^k \subset r_k$(see Figure 1(left)), the support of r is the proportion of objects in $\mathcal{I}$ that satisfy the antecedent of the rule, Liu [2000]. $Sup(r) = \frac{card\{i \in \mathcal{C} \ : \ x_{ik} \in I_s^k\}}{n}$
- Relative covering $(CovR)$: Given a rule, the relative covering is the proportion of class C that satisfy the rule. $CovR(r) = \frac{card\{i \in C \ : \ x_{ik} \in I_s^k\}}{card\{C\}} * 100$
- Confidence $(Conf)$: Given a rule, the confidence of r is the proportion of objects in $\mathcal{I}$ that satisfy the antecedent of the rule and belong to C, Liu [2000]. $Conf(r) = \frac{card\{i \in C \ : \ x_{ik} \in I_s^k\}}{card\{ \ x_{ik} \in I_s^k\}}$
- *B*oxplot based discretization (BbD), see Pérez-Bonilla and Gibert [2007]), as an efficient way of transforming a numerical variable into a qualitative one in such a way that the resulting qualitative variable maximizes the association with the reference partition. Basically the cut points are determine by the minimum and maximum values that the numerical variable take in a very groups induced by the categorical one.
- The methodology *boxplot based induction rules (BbIR)* is presented in Pérez-Bonilla and Gibert [2007]. It is a method for inducing probabilistic rules ($r : x_{ik} \in I_s^k \xrightarrow{p_{sc}} C$, $p_{sc} \in [0, 1]$ is a degree of certainty of r) with a minimum number of attributes in the antecedent, based on the *BbD* of X_k.

The standard input of a clustering algorithm is a data matrix with the values of K variables $X_1 \dots X_K$ (numerical or not) observed over a set $\mathcal{I} = \{1, \dots n\}$ of individuals. Variables are in columns, while individuals in rows. Cells contain the value (x_{ik}), taken by individual $i \in \mathcal{I}$ for variable $X_k, (k = 1 : K)$. The set of values of X_k is named $\mathcal{D}^k = \{c_1^k, c_2^k, ..., c_s^k\}$ for categorical variables and $D^k = r_k$ for numerical ones, being r_k = $[minX_k$,$maxX_k]$ the range of X_K. A partition in ξ classes of $\mathcal{I}$ is denoted by $\mathcal{P}_\xi = \{C_1, ..., C_\xi\}$, and $\tau = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, ..., \mathcal{P}_n\}$ is an indexed hierarchy of $\mathcal{I}$. Finally, $\mathcal{P}_2 = \{C_1, C_2\}$ is a binary partition of $\mathcal{I}$. Usually, τ is the result of a *hierarchical clustering* over $\mathcal{I}$, and it can be represented in a graphical way as a *dendrogram* (or hierarchical tree, see figure 1, Pérez-Bonilla et al. [2007]).

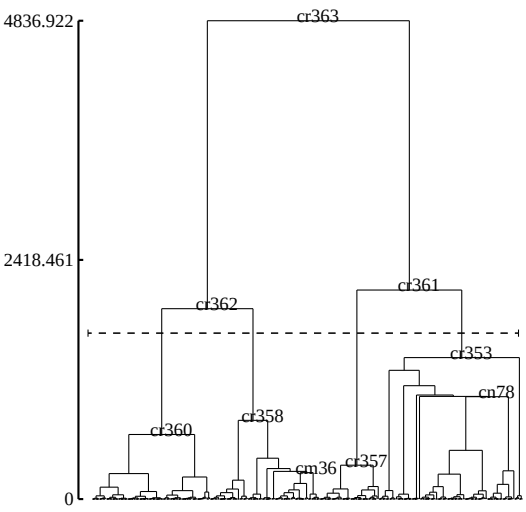

Figure 1: *Left:* Notation; *Right:* Hierarchical tree [$\tau_{Lj3,R2}^{EnW,G}$] (Pérez-Bonilla et al. [2007]) .

3 THE METHODOLOGY

CCEC takes advantage of the existence of τ, and 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 $\mathcal{P}_{\xi+1}$. Binary hierarchical structure will be used by *CCEC* to discover particularities of the final classes step by step also in hierarchical way. The *CCEC*, Pérez-Bonilla and Gibert [2007], allows generation of automatic conceptual interpretations of a given partition $\mathcal{P} \in \tau$. The steps to be followed are described below. The application of CCEC to the WWTP is illustrated in §5:

1. Cut the tree at highest level (make $\xi = 2$ and consider $\mathcal{P}_2 = \{C_1, C_2\}$).
2. Use *BbD* (Pérez-Bonilla and Gibert [2007]), to find (total or partial) characteristic values for numerical variables, Gibert et al. [1998].
3. Use *BbIR*, to induce a knowledge Base describing both classes.
4. For classes in $\mathcal{P}_2$, determine concepts A_1^{ξ,X_k} : "$[X_k \in I_s^k]$", $A_2^{\xi,X_k} : \neg A_1^{\xi,X_k}$ associated to C_1, C_2, by taking the intervals provided by a totally characteristic variable or the partial one with greater relative covering and $p_{sc} = 1$.
5. Go down one level in the tree, by making $\xi = \xi + 1$ and so considering $\mathcal{P}^{\xi+1}$. As said before $\mathcal{P}^{\xi+1}$ is *embedded* in $\mathcal{P}^\xi$ in such a way that there is a class of $\mathcal{P}^\xi$ splitting in two

new classes of $\mathcal{P}^{\xi+1}$, namely $C_i^{\xi+1}$ and $C_j^{\xi+1}$ and all other classes $C_q^{\xi+1}, q \neq i, j$, are common to both partitions and $C_q^{\xi+1} = C_q^{\xi} \ \forall q \neq i, j$.
Since in the previous step $C_i^{\xi+1} \cup C_j^{\xi+1}$ were conceptually separated from the rest, at this point it is only necessary to find the variables which separate (or distinguishes) $C_i^{\xi+1}$ from $C_j^{\xi+1}$, by repeating steps 2-4. Suppose $B_i^{\xi+1,X_k}$ and $B_j^{\xi+1}$ the concepts induced from $C_i^{\xi+1}$ and $C_j^{\xi+1}$, in the step $\xi + 1$.

6. Integrate the extracted knowledge of the iteration $\xi + 1$ with that of the iteration ξ, by determining the compound concepts finally associated to the elements of $\mathcal{P}_{\xi+1}$. The concepts for the classes of $\mathcal{P}_{\xi+1}$ will be: $A_q^{\xi+1,X_k} = A_q^{\xi,X_k}$, $A_i^{\xi+1,X_k} = \neg A_q^{\xi,X_k} \wedge B_i^{\xi+1,X_k}$ and $A_j^{\xi+1,X_k} = \neg A_q^{\xi,X_k} \wedge B_j^{\xi+1,X_k}$
7. Make $\xi = \xi + 1$, and return to the step 5 repeating until $\mathcal{P}_\xi = \mathcal{P}$.

4 CASE STUDY

A case study in this paper was the pilot plant, which is located in Domale-Kamnik wastewater treatment plant in Slovenia. A scheme of the pilot plant with sensors and actuators is shown in Figure 2. In the pilot plant the moving bed biofilm reactor (MBBR) technology is tested for the purpose of upgrading the whole plant for nitrification and denitrification. The pilot plant with the volume of 1125 m3 consists of two anoxic and two aerobic tanks that are filled with the plastic carriers on which the biomass develops, a fifth tank, which is a dead zone without plastic carriers and a settler. The total air flow to the both aerobic tanks can be on-line manipulated in such a way that oxygen concentration in the first aerobic tank is controlled at the desired value. The wastewater rich with nitrate is recycled with the constant flow rate from the fifth tank back to the first tank. The influent to the pilot plant is wastewater after mechanical treatment, which is pumped to the pilot plant. The inflow is kept constant to fix the hydraulic retention time. The influent flow rate can be adjusted manually to observe the plant performance at different hydraulic retention times.

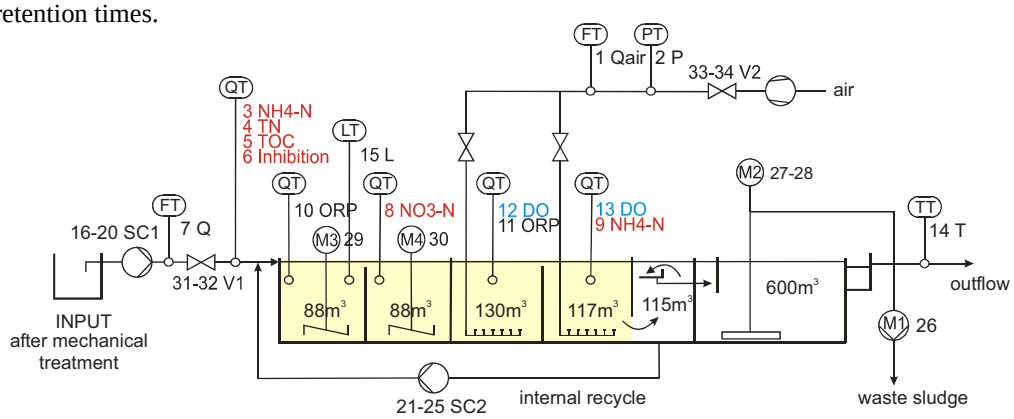

Figure 2: MBBR (Moving Bed Biofilm Reactor) pilot plant with sensors and actuators.

The database that was used in this study consists of 365 daily averaged observations that were taken from the 1st of June 2005 to the 31th of May 2006. Every observation includes measurements of the 16 variables that are relevant for the operation of the pilot plant. The variables that were measured are:

- NH4-infuent: ammonia concentration at the influent of the pilot plant(pp) (3 in the Fig. 2).
- Q-influent: wastewater influent flow rate of the pp (7 in the Fig. 2).
- TN-influent: concentration of the total nitrogen at the influent of the pp (4 in the Fig. 2).
- TOC-influent: total organic carbon concentration at the influent of the pp (5 in the Fig. 2).
- Nitritox-influent: measurement of the inhibition at the influent of the pp (6 in the Fig. 2).
- h-wastewater: height of the wastewater in the tank (no in the Fig. 2).
- O2-1aerobic: dissolved oxygen concentration in the 1st aerobic tank (3rd tank) (12-Fig. 2).
- Valve-air: openness of the air valve (between 0 - 100%), highly related with Q-air (V2 in the Fig. 2).
- Q-air: total air flow that is dosed in both aerobic tanks (1 in the Fig. 2).

- NH4-2aerobic: ammonia concentration in the second aerobic tank (9 in the Fig. 2).
- O2-2aerobic: dissolved oxygen concentration in the 2nd aerobic tank (4th tank)(13-Fig. 2).
- TN-effluent: concentration of the total nitrogen at the effluent of the pp (no in the Fig. 2).
- Temp-wastewater: temperature of the wastewater (14 in the Fig. 2).
- TOC-effluent: total organic carbon concentration at the effluent of the pp (no in the Fig. 2).
- Freq-rec: frequency of the internal recycle flow rate meter (no in the Fig. 2).
- FR1-DOTOK-20s (Hz): frequency of the motor that pums the wastewater into the pilot plant. This frequency is highly correlated with Q-inffluent.

The data base was clustered in a previous work by using clustering based on rules proposed by Gibert [1996] using the following Knowledge Base:

$KB = \{((and(>= (NH4-2aerobic)10.0)(> (TN-effluent)18.0)) \rightarrow Mmonia),$
$((and(< \quad (NH4-2aerobic)10.0)(> (TN-effluent)18.0)) \rightarrow Nitrogen)\}$

with 38 objects in class Mmonia, 80 objects in class Nitrogen and Residual class with 247 objects. The results are presented in Pérez-Bonilla et al. [2007]. The hierarchical tree of Figure 1(right) was produced and a $\mathcal{P}_4 = \{Cr_{353}, Cr_{357}, Cr_{358}, Cr_{360}\}$ is obtained. Figure 3 contains the *class panel graph*, Gibert et al. [2005] of the 16 variables regarding the partition $\mathcal{P}_4$ where the *multiple boxplot*, Tukey [1977] of variables for each class are displayed. 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 guarantees the *distinguishability* of classes. The result is a 4-class partition $\mathcal{P}_4$.

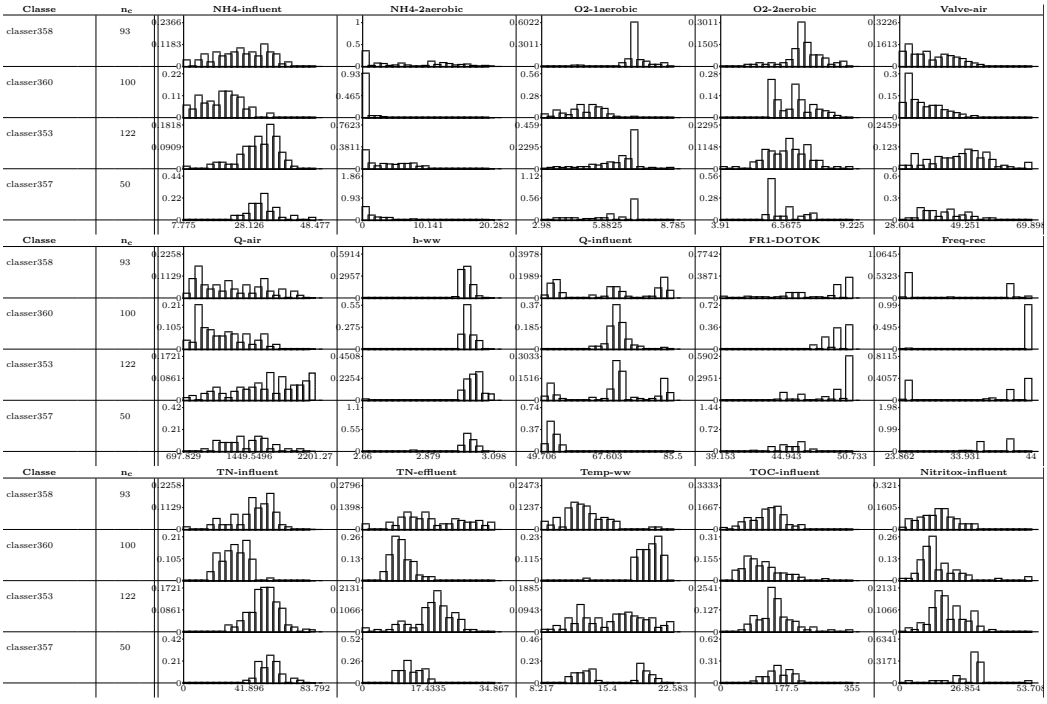

Figure 3: Class panel graph of $\mathcal{P}_4 = \{Cr_{353}, Cr_{357}, Cr_{358}, Cr_{360}\}$.

From the class panel graph, experts provided the following interpretation:

- Cr_{353}, represents the plant operation under the high load. In this case influent nitrogen concentrations are high and also influent flow rate is quite high as well. Even though the oxygen concentration in the aerobic tanks are high this can not decrease the effluent nitrogen concentrations. It means that when the plant is overloaded high effluent concentrations at the effluent of the plant can be expected.
- Cr_{357}, represents the situation when the influent flow rate is low, that is, when the hydraulic retention time of the plant is high. In this case we get quite low effluent nitrogen concentrations if of course oxygen concentration in the aerobic tank is high enough. It means when the influent flow rate to the plant is low the effluent concentrations of the plant can be obtained at the low level if the oxygen concentration in the aerobic tanks is high.

- Cr_{358}, explains the situation when the wastewater temperature is low. In this case nitrogen removal efficiency of the plant is rather low. This is so because microorganisms in the tanks don't work so intensively in cold conditions and therefore higher concentrations at the effluent of the plant can be expected.
- Cr_{360}, shows the situation when the wastewater temperature is high. In warmer conditions the microorganisms in the plant work faster, so the effluent nitrogen concentrations can be low even when the oxygen concentrations in the aerobic tanks are quite low.

5 RESULTS

Application of the methodology presented in §3 produced the following results.

The 2-class Partition. For the presented data the following 2-class partition $\mathcal{P}_2 = \{C_{361}, C_{362}\}$ is obtained. As stated in §2, I^k is built using *Boxplot based discretization (BbD)* for all variables. Next, *BbIR* is used to generate all the rules induced for $\mathcal{P}_2$, by taking the intervals provided by a totally characteristic variable or the partial one with greater relative covering and $p_{sc} = 1$. Here, TN-influent has the greatest relative covering (22,80%), see Table 1.

The following association concept can be done.

- $A^{2,TN-influent}_{Cr_{362}}$= "$x_{TN-influent,i} \in [0.0, 28.792)$" is associated with Cr_{362}
- $A^{2,TN-influent}_{Cr_{361}}$=$\neg A^{2,TN-influent}_{Cr_{362}}$= "$x_{TN-influent,i} \in [28.79, 83.79]$" associated with Cr_{361}

Or, in other words:

- Class Cr_{361}, "Not low Concentration of the total nitrogen at the influent".
- Class Cr_{362}, "Low Concentration of the total nitrogen at the influent".

The next step is to go one level down the tree.

Table 1: Summary of Knowledge Base for Cr_{361} and Cr_{361} from $P2^{EnW,G}_{Lj3,R2}$

Concep	Knowledge Base Cr_{361} (172 days) and Cr_{362} (193 days)	Cov	$CovR$
$A^{2,NH4-influent}_{C_{361}}$	$r^{NH4-influent}_{3,Cr_{361}}$: $x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} Cr_{361}$	5	2.91%
$A^{2,O2-1aerobic}_{C_{361}}$	$r^{O2-1aerobic}_{3,Cr_{361}}$: $x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{1.0} Cr_{361}$	1	0.58%
$A^{2,O2-2aerobic}_{C_{361}}$	$r^{O2-2aerobic}_{1,Cr_{361}}$: $x_{O2-2aerobic,i} \in [3.91, 4.94) \xrightarrow{1.0} Cr_{361}$	2	1.16%
$A^{2,Valve-air}_{C_{361}}$	$r^{Valve-air}_{3,Cr_{361}}$: $x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr_{361}$	**28**	**16.28%**
$A^{2,Q-air}_{C_{361}}$	$r^{Q-air}_{3,Cr_{361}}$: $x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{1.0} Cr_{361}$	27	15.70%
$A^{2,h-ww}_{C_{361}}$	$r^{h-ww}_{3,Cr_{361}}$: $x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{1.0} Cr_{361}$	16	9.30%
$A^{2,Q-influent}_{C_{361}}$	$r^{Q-influent}_{1,Cr_{361}}$: $x_{Q-influent,i} \in [49.706, 50.99) \xrightarrow{1.0} Cr_{361}$	6	3.49%
$A^{2,Freq-rec}_{C_{361}}$	$r^{Freq-rec}_{3,Cr_{361}}$: $x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{1.0} Cr_{361}$	3	1.74%
$A^{2,TN-influent}_{C_{361}}$	$r^{TN-influent}_{3,Cr_{361}}$: $x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{1.0} Cr_{361}$	9	5.23%
$A^{2,Temp-ww}_{C_{361}}$	$r^{Temp-ww}_{3,Cr_{361}}$: $x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{1.0} Cr_{361}$	5	2.91%
$A^{2,TOC-influent}_{C_{361}}$	$r^{TOC-influent}_{3,Cr_{361}}$: $x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{1.0} Cr_{361}$	20	11.63%
$A^{2,TOC-effluent}_{C_{361}}$	$r^{TOC-effluent}_{3,Cr_{361}}$: $x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{1.0} Cr_{361}$	10	5.81%
$A^{2,NH4-influent}_{C_{362}}$	$r^{NH4-influent}_{1,Cr_{362}}$: $x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{1.0} Cr_{362}$	3	1.55%
$A^{2,NH4-2aerobic}_{C_{362}}$	$r^{NH4-2aerobic}_{3,Cr_{362}}$: $x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{1.0} Cr_{362}$	38	19.69%
$A^{2,O2-1aerobic}_{C_{362}}$	$r^{O2-1aerobic}_{1,Cr_{362}}$: $x_{O2-1aerobic,i} \in [2.98, 3.297) \xrightarrow{1.0} Cr_{362}$	4	2.07%
$A^{2,Valve-air}_{C_{362}}$	$r^{Valve-air}_{1,Cr_{362}}$: $x_{Valve-air,i} \in [28.604, 28.934) \xrightarrow{1.0} Cr_{362}$	5	2.59%
$A^{2,Q-air}_{C_{362}}$	$r^{Q-air}_{1,Cr_{362}}$: $x_{Q-air,i} \in [697.829, 739.819) \xrightarrow{1.0} Cr_{362}$	2	1.04%
$A^{2,Q-influent}_{C_{362}}$	$r^{Q-influent}_{3,Cr_{362}}$: $x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{1.0} Cr_{362}$	1	0.52%
$A^{2,FR1-DOTOK}_{C_{362}}$	$r^{FR1-DOTOK}_{1,Cr_{362}}$: $x_{FR1-DOTOK,i} \in [39.153, 42.276) \xrightarrow{1.0} Cr_{362}$	5	2.59%
$A^{2,TN-influent}_{C_{362}}$	$r^{TN-influent}_{1,Cr_{362}}$: $x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} Cr_{362}$	**44**	**22.80%**
$A^{2,TN-efluent}_{C_{362}}$	$r^{TN-effluent}_{3,Cr_{362}}$: $x_{TN-effluent,i} \in (28.933, 34.867] \xrightarrow{1.0} Cr_{362}$	14	7.25%
$A^{2,TOC-influent}_{C_{362}}$	$r^{TOC-influent}_{1,Cr_{362}}$: $x_{TOC-influent,i} \in [0.0, 63.22) \xrightarrow{1.0} Cr_{362}$	20	10.36%
$A^{2,Nitritox-influent}_{C_{362}}$	$r^{Nitritox-influent}_{1,Cr_{362}}$: $x_{Nitritox-influent,i} \in [0.0, 3.833) \xrightarrow{1.0} Cr_{362}$	4	2.07%
$A^{2,TOC-effluent}_{C_{362}}$	$r^{TOC-effluent}_{1,Cr_{362}}$: $x_{TOC-effluent,i} \in [0.0, 2.014) \xrightarrow{1.0} Cr_{362}$	1	0.52%

The 3-class partition. Take $\mathcal{P}_3 = \{Cr_{362}, Cr_{353}, Cr_{357}\}$ and first identify correspondences between the classes of $\mathcal{P}_2 = \{Cr_{361}, Cr_{362}\}$ and $\mathcal{P}_3 = \{Cr_{362}, Cr_{353}, Cr_{357}\}$. Class Cr_{361} splits into Cr_{353} (referred as $Cr_i^{\xi+1}$ in methodology), Cr_{357} (referred as $Cr_j^{\xi+1}$ in methodology) while Cr_{362} (Cr_{362} is referred as CR_q^{ξ} in methodology) remains in both $\mathcal{P}_2$ and $\mathcal{P}_3$. From the previous iteration it is already known that there is a common characteristic of both classes, Cr_{357} and Cr_{353}, (TN-influent is high) which distinguishes them from Cr_{362} (with low TN-influent). Thus, it is only required to find the separation between Cr_{353} and Cr_{357}. A similar procedure as was used in previous step for separating CR_{362} and Cr_{361} was used with all the variables regarding Cr_{353} and Cr_{357}. Here Q-influent has the greatest relative covering of 81,97%. Combining with the results of previous iteration of *CCEC* leads to the following interpretation of $\mathcal{P}_3$:

-Cr_{362} is such that, $A^{2,TN-influent}_{Cr_{362}}$= "$x_{TN-influent,i} \in [0.0, 28.792)$"
-Cr_{353} is such that, $\neg A^{2,TN-influent}_{Cr_{362}} \wedge B^{3,Q-influent}_{Cr_{353}}$
"$x_{TN-influent,i} \in [28.792, 83.792]$" $\wedge$ "$x_{Q-influent,i} \in (55.666, 85.092]$"
-Cr_{357} is such that, $\neg A^{2,TN-influent}_{Cr_{362}} \wedge \neg B^{3,Q-influent}_{Cr_{353}}$
"$x_{TN-influent,i} \in [28.792, 83.792]$" $\wedge$ "$x_{Q-influent,i} \in (49.706, 55.666]$"

Degree of certainty should be associated depending on the probabilities of the corresponding generated concepts.

The final partition. The process would continue separating the Cr_{360} and Cr_{358} of the partition $\mathcal{P}_4$ (see figure 1(right)), which are the subdivision of Cr_{362}. Here, a Temp-ww with a relative covering of 66,67%, is chosen. Similarly, the interpretation of $\mathcal{P}_4$, which is the final partition obtained with the Knowledge Base with certain rules for Cr_{360} and Cr_{358} is as follows:

– Cr_{353} is such that "$x_{TN-influent,i} \in [28.792, 83.792]$" $\wedge$ "$x_{Q-influent,i} \in (55.666, 85.092]$"
"Non low Concentration of the total nitrogen at the influent and high values of Wastewater influent flow rate".
– Cr_{357} is such that "$x_{TN-influent,i} \in [28.792, 83.792]$" $\wedge$ "$x_{Q-influent,i} \in [49.706, 55.666]$
"Non low Concentration of the total nitrogen at the influent and not high values of Wastewater influent flow rate".
– Cr_{358} is such that "$x_{TN-influent,i} \in [0.0, 28.792)$" $\wedge$ "$x_{Temp-ww,i} \in [8.472, 13.327)$"
"Low Concentration of the nitrogen and low values of water's temperature."
– Cr_{360} is such that "$x_{TN-influent,i} \in [0.0, 28.792)$" $\wedge$ "$x_{Temp-ww,i} \in [13.327, 21.896]$"
"Low Concentration of the nitrogen and non low values of temperature".

This set of concepts can, in fact, be considered as a domain model which can support later decision for the treatment that is applied to a new day, provided that a standard treatment is previously associated to every class by experts. In this association the possibility of easily interpreting the classes is critical as well as the knowledge from the experts for easily understanding the meaning of the classes. In this sense the proposed method provides simple and short rules which use to be easier to handle than those provided by other induction rules algorithms.

Evaluating Confidence and Support terms and comparing them with the class interpretation provided by the expert two modifications of the method are assessed in this work. First includes the best (the greatest relative covering) rule for every class in the description and seconds uses a closed-world assumption.

Best local concept: Here Valve-air is chosen for Cr_{361} and TN-influent for Cr_{362}, see Table 1, Q-influent is chosen for the Cr_{353} ($CovR = 81,97\%$) and FR1-DOTOK ($CovR = 16,00\%$) for Cr_{357} and Temp-ww is chosen for the Cr_{360} ($CovR = 29,00\%$) and Cr_{358} ($CovR = 66,67\%$). The final interpretation of 4 classes is the following:

– Cr_{353} is such that "$x_{Valve-air,i} \in (54.77, 69.898]$" $\wedge$ "$x_{Q-influent,i} \in (55.666, 85.092]$"
High percentages of the air valve openess and high values of Wastewater influent flow rate.
– Cr_{357} is such that "$x_{Valve-air,i} \in (54.77, 69.898]$" $\wedge$ "$x_{FR1-DOT,i} \in [42.276, 44.167)$"
High percentages of the air valve openess and low values of frequency of the influent flow rate meter.

– Cr_{358} is such that "$x_{TN-influent,i} \in [0.0, 28.792) \wedge$ "$x_{Temp-ww,i} \in [8.472, 13.327)$"
Low values of the total nitrogen concentration at the influent and low values of the wastewater temperature.
– Cr_{360} is such that "$x_{TN-influent,i} \in [0.0, 28.792) \wedge$ "$x_{Temp-ww,i} \in (20.928, 21.896]$"
Low values of the total nitrogen concentration at the influent and high values of the wastewater temperature.

Best local concept and partial Close-World Assumption: Includes the same variables as the first modification together with the complementary concepts of the other class except when the best variable for the two classes is the same in iteration. In this case the original concepts is kept (this criteria behaves better than CCEC, if does the comparison with the concept those expert).

- Cr_{353} is such that ("$x_{TN-influent,i} \in [28.792, 83.792]$" $\vee$ "$x_{Valve-air,i} \in (54.77, 69.898]$") $\wedge$
("$x_{Q-influent,i} \in (55.666, 85.092]$" $\vee$ "$x_{FR1-DOTOK,i} \in [44.167, 50.7]$")
"High-medium of the total nitrogen concentration at the influent or high percentages of the air valve openess and high values of Wastewater influent flow rate or high-medium values of the influent motor pump frequency"
- Cr_{357} is such that ("$x_{TN-influent,i} \in [28.792, 83.792]$" $\vee$ "$x_{Valve-air,i} \in (54.77, 69.898]$") $\wedge$
("$x_{Q-influent,i} \in [49.706, 55.666]$ " $\vee$ "$x_{FR1-DOTOK,i} \in [42.276, 44.167)$")
"High-medium of the total nitrogen concentration at the influent or high percentages of the air valve openess and low-medium values of Wastewater influent flow rate or low values of the influent motor pump frequency"
- Cr_{358} is such that "($x_{TN-influent,i} \in [0.0, 28.792) \vee$ " $x_{Valve-air,i} \in [28.604, 54.777]) \wedge$
"$x_{Temp-ww,i} \in [8.472, 13.327)$"
"Low of the total nitrogen concentration at the influent or low-medium percentages of the air valve openess and low values of the wastewater temperature."
- Cr_{360} is such that "$x_{TN-influent,i} \in [0.0, 28.792) \vee$ "$x_{Valve-air,i} \in [28.604, 54.777]) \wedge$
"$x_{Temp-ww,i} \in (20.928, 21.896]$"
"Low of the total nitrogen concentration at the influent or low-medium percentages of the air valve openess and high values of the wastewater temperature."

Three approaches generate different concepts for each class. Support and confidence factors together with average confidences and global support of concepts, are shows in Table 2.

Table 2: Comparison among the 3 proposals

	CCEC Best global rule		Best local rule		Best local rule+CWA global		partial	
Ruler	$Conf.$	$Supp.$	$Conf.$	$SupP.$	$Conf.$	$Supp.$	$Conf.$	$Supp.$
$\mathcal{R}_{Cr_{353}}$	45,00%	27,12%	100,0%	7,40%	39,48%	32,42 %	39,48%	33,42%
$\mathcal{R}_{Cr_{357}}$	46,94%	12,60%	100,0%	0,27%	47,52%	13,15%	47,52%	13,15%
$\mathcal{R}_{Cr_{358}}$	100,0%	1,64%	100,0%	1,64%	54,79%	16,99 %	54,87%	16,99%
$\mathcal{R}_{Cr_{360}}$	86,84%	9,04%	100,0%	0,82%	45,00%	27,12 %	82,86%	7,95%
Average	69,695%		100%		46,697%		56,18%	
Global Supp.		50,41%		10,13%		90,68%		71,51%

6 CONCLUSIONS AND FUTURE WORK

In this paper a methodology to generate automatic conceptual interpretations of a group of classes is presented. Concepts associated with classes are built taking advantage of hierarchical structure of the underlying clustering. The *Conceptual characterization by embedded conditioning CCEC*, Gibert and Pérez-Bonilla [2006], Pérez-Bonilla and Gibert [2007], is a quick and effective method that generates a conceptual model of the domain, which will be of great support to the later decision making based on a combination of *BbD* and an interactive combination of concepts upon hierarchical subdivisions of the domain. This is a preliminary proposal that has been applied with success to real data coming from a WWTP. Benefits of this proposal are specially interesting in the interpretation of partitions with a large number of classes. Automatic generation of interpretations

cover the important goal of KDD of describing the domain Fayyad [1996.]. However, in this proposal a direct connection between the generated concepts and the automatic rules generation allows direct construction of a decision model for the later class prediction. As a matter of a fact, automatic production of probabilistic or fuzzy classification rules regarding concepts provided by *CCEC* is direct, as discussed in Gibert and Pérez-Bonilla [2005]. By associating an appropriate characteristic to every class a model for how to operate the wastewater plant based on a concrete day upon a reduced number variables is obtained together with an estimation of the risk associated to that decision (which is related with the certainty of the rule). In this work three different criteria for deciding which variable is kept at every iteration are assessed. The one that gives the most similar interpretation to those expert is the Best local rule +partialCWA, which from a technical point of view also seems to represent the more equilibrated option with high values in confidence and support factors. In future work we will study how to propagate uncertainty from one iteration to the next needs to be analyzed in depth (here rules with $p_{sc} = 1$ is used). The idea is to use an approach which avoids the explicit construction of all the concepts to evaluate their relative coverage. Comparison of rules produced with *CCEC* and with other methods have been presented in Gibert et al. [2006]. Finally a modification of the methodology is required to guarantee that important variables at the output are also included in the final description. In long term proposal could be extended to keep more the one variable per iteration.

ACKNOWLEDGMENTS

We would like to thank the staff of the Domale-Kamnik WWTP for providing us with the data from the pilot plant and the Proyect TIN 2004-01368.

REFERENCES

Fayyad, U. *From Data Mining to Knowledge Discovery: An overview*. 1996.

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

Gibert, K., T. Aluja, and U. Cortés. Knowledge Discovery with Clustering Based on Rules. Interpreting results. LNAI v. 1510, pages 83–92. Springer, 1998.

Gibert, K., R. Nonell, J. M. Velarde, et.al. Knowledge discovery with clustering: impact of metrics and reporting phase by using klass. *Neural Network World*, 4/05:319–326, 2005.

Gibert, K. and A. Pérez-Bonilla. Taking advantage of the hierarchical structure of a clustering for automatic generation of classification interpretations. In *Fuzzy sets in learning and data mining.* , pages 524–529, Barcelona, España, septiembre 2005. EUSFLAT.

Gibert, K. and A. Pérez-Bonilla. Towards automatic generation of interpretation as a tool for modelling decisions. In *Proceedings of III International Conference on Modeling Decisions for Artificial Intelligence*, pages 515–524, Tarragona, Abril 2006. MDAI.

Gibert, K., A. Pérez-Bonilla, and G. Rodriguez. A Comparative Analysis of different techniques for class interpretation supporting tools. In *In Procs. Nov Congrs Catal d'Intelligncia Artificial.*, pp. 37–46,IOS press. Perpignan, France, noviembre 2006. CCIA06.

Gordon, A. D. Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, pages V.18: 561–581, 1994.

Liu, B. Hsu, W. C. S. M. Y. Analyzing the subjective interestigness of association rules. *IEEE Intelligent Systems*, pages 47–55, 2000.

Pérez-Bonilla, A. and K. Gibert. Towards automatic generation of conceptual interpretation of clustering. In *Progress in Pattern recognition, Image analysis and Application.*, volume 4756 of *Lecture Notes in Computer Science*, pp. 653–663, Springer-Verlag. Valparaiso-Chile, 2007.

Pérez-Bonilla, A., K. Gibert, and D. Vrecko. Domale-kamnik wastewater treatment plant (ljubljana - slovenia). clustering and induction knowledge base. Research DR 2007/09, Dep. Estadística e Investigación Operativa. UPC, Barcelona, España, Diciembre 2007.

Tukey, J. *Exploratory Data Analysis*. Addison-Wesley, 1977.