International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Automatic interpretation of classes for improving decision support

K. Gibert [a], **A. Pérez-Bonilla**[b]

[a]*Knowledge Engineering and Machine Learning group.*
Universitat Politècnica de Catalunya, Edif. C5, Jordi Girona 1-3, 08034 Barcelona, SPAIN
(karina.gibert@upc.edu)

[b]*Department of Statistics and Operations Research. Universitat Politècnica de Catalunya, Edif. C5, Jordi Girona 1-3, 08034 Barcelona, SPAIN)*

Abstract: More and more, the analysis of clustering results becomes difficult as the number of variables considered increases, and the number of classes is not low. Sometimes concept induction methods are used to associate concepts to every class and use to be expressed as boolean expressions, easy to understand and supposedly providing good support to decision making. It has been seen that most of the concept induction algorithms priorize the compacity of the final expressions, as well as their predictive power. However, for descriptive purposes, when the meaning of classes has to be recognized and understood by the expert, this is not the best approach, since compacity directly implies elimination of redundancies or strong associations, while comprehension of the class is mainly based in understanding how variables interact among them inside the class. Here a method to induce conceptual descriptions of classes is proposed, providing non-minimal descriptions of the classes, but richer ones including the characteristics that distinguishes a class from the others, in such a way that expert can easily recognize the essence of the class, and conceptualize it on the bases of local interactions among all variables observed inside every class. This kind of interpretations provide an excellent support for later decision support systems.

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* Volle [1985.], 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 Hand [1996]. 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 modeled and it is oriented to develop, in the long term, intelligent decision support systems.

The presented proposal integrates different findings from a series of previous works: 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 et al. [1998], Gibert [1996]. In the present work, for the first time, the whole proposal, named Conceptual Characterization by Embedded Conditioning (CCEC), is applied to a real environmental data set and different possibilities for integrating knowledge from one iteration to the following one are exhaustively compared and evaluated. For the first time this work presents a depth analysis of the quality of the solutions provided by 5 different strategies, both considering structural quality criteria, as confidence or support of the provided descriptions as well as more semantic criteria, as the proximity towards the descriptions provided by the experts.

This paper is organized as follows: After the introduction, the methodology is presented in §2. §3 introduces the waste water treatment plant (WWTP) of the case study and the data used. Results of applying *CCEC* to the data described are given in §4. Finally in §5 the conclusions and the future work are addressed.

2 FORMAL FRAME

The standard input of a clustering algorithm is a data matrix with the values of K variables $X_1 \ldots X_K$ (numerical or not) observed over a set $\mathcal{I} = \{1, \ldots n\}$ 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 an horizontal cut of the corresponding *dendrogram* (or hierarchical tree, see Figure 1, Pérez-Bonilla et al. [2008]).

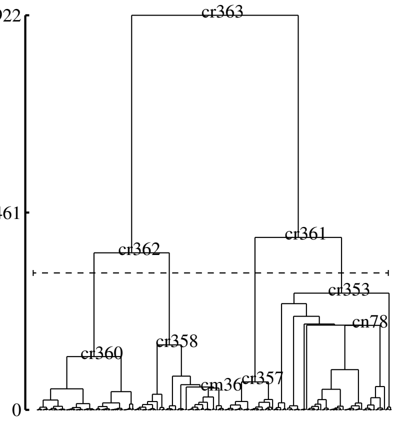

Figure 1: Dendrogram $[\tau^{EnW,G}_{Lj3,R2}]$.

CCEC is a methodology globally described in Pérez-Bonilla et al. [2008] that takes advantage of the existence of τ to generate conceptual interpretations of a of a given partition $\mathcal{P} \in \tau$ in terms of formal descriptions. CCEC uses the property of all binary hierarchical structure that $\mathcal{P}_{\xi+1}$ has the same classes of $\mathcal{P}_\xi$ except one, which splits in two subclasses in $\mathcal{P}_{\xi+1}$. The binary hierarchical structure represented in τ is used in *CCEC* to discover particularities of the final classes step by step by analyzing the hierarchy top-down. It uses *B*oxplot based discretization (BbD), see Gibert and Pérez-Bonilla [2006]), as an efficient way of transforming all numerical variable into qualitative ones in such a way that every resulting qualitative variable maximizes the association with the reference partition. See Gibert and Pérez-Bonilla [2006] for details. Briefly, main idea is to use as cut-points the extreme values (minimum and maximum) that the numerical variable locally takes in every class of $\mathcal{P}$. *BbD* is the kernel of *Boxplot based induction rules (BbIR)* (presented in Pérez-Bonilla and Gibert [2007]). It is a method for inducing probabilistic rules ($r : x_{ik} \in I_s^k \xrightarrow{p_{sc}} C$, being $p_{sc} \in [0, 1]$ the certainty degree of r). The produced rules have a minimum number of attributes in the antecedent, and those are formalized on the basis of the intervals induced by *BbD* for every variable. The CCEC methodology was formalized in Pérez-Bonilla et al. [2008]. Here an algorithmic version is presented:

1. Consider the top of the tree: $\xi = 1$; $\mathcal{P}_1 = \mathcal{I}$; $\mathcal{A}_{\mathcal{P}_1} = \{A^1 : true\}$
2. Go down one level in the tree, by making $\xi = \xi + 1$ and so considering the new $\mathcal{P}^\xi$. Being τ an indexed hierarchy, $\mathcal{P}^\xi$ is *embedded* in $\mathcal{P}^{\xi-1}$ in such a way that there is a single class of $\mathcal{P}^{\xi-1}$, namely $C_t^{\xi-1}$, splitting in two new classes of $\mathcal{P}^\xi$, namely C_i^ξ and C_j^ξ and all other classes $C_q^\xi, q \neq i, j$, are common to both partitions and $C_q^\xi = C_q^{\xi-1} \ \ \forall q \neq i, j$. Consider the restricted partition $\mathcal{P}_\xi^* = \{C_i^\xi, C_j^\xi\}$. It holds that $\mathcal{P}_\xi^* \subset \mathcal{P}^\xi$ and when $\xi = 2$, $\mathcal{P}_\xi^* = \mathcal{P}^\xi$. As in previous iteration the class $C_t^{\xi-1} = \{C_i^\xi \wedge C_j^\xi\}$ was already distinguished from the rest by proper concept, it is enough to find distinction between C_i^ξ and C_j^ξ.

3. Use *BbD* (Gibert and Pérez-Bonilla [2006]), to find (total or partial) characteristic values regarding $\mathcal{P}^*_\xi$ Gibert et al. [1998] for all numerical variables.
4. Use *BbIR*, to induce a knowledge base $\mathcal{R}(\mathcal{P}^*_\xi)$ describing both classes $\{C_i^\xi, C_j^\xi\}$.
5. Search the best rule for each class of the restricted partition $P^*_\xi = \{C_i^\xi, C_j^\xi\}$. In the next section several criteria are presented to determine them. Name $A_i^{*\xi}$ and $A_j^{*\xi}$ the antecedents of the rules selected for C_i^ξ and C_j^ξ respectively.
6. Integrate $A_i^{*\xi}$ and $A_j^{*\xi}$ with the father's concept from previous iteration. Compound concepts are associated to C_i^ξ and C_j^ξ:

$$A_i^\xi = A_t^{\xi-1} \wedge A_i^{*\xi} \quad ; \quad A_j^\xi = A_t^{\xi-1} \wedge A_j^{*\xi} \tag{1}$$

Description of both C_i^ξ and C_j^ξ inherits the properties of the father class $C_t^{\xi-1}$.
7. Build the concepts system:

$$\mathcal{A}_{\mathcal{P}_\xi} = \mathcal{A}_{\mathcal{P}_{\xi-1}} \backslash \{C_t : \mathcal{A}_t\} \cup \{C_i^\xi : A_i^\xi, C_j^\xi : A_j^\xi\}$$

8. Go down one level in the tree, by making $\xi = \xi + 1$ and so considering $\mathcal{P}^{\xi+1}$. Return to **step 2** and repeat until $\mathcal{P}^\xi = \mathcal{P}$, $\mathcal{P}$ target partition to be interpreted.
9. Finally, $\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \; \forall C \in \mathcal{P}_\xi\}$ and also, the concepts system can be associated to a rules system $\mathcal{R}(\mathcal{P}^\xi) = \{r \; tq \; r : A \xrightarrow{p(r)} C \; \forall C \in \mathcal{P}_\xi\}$.

The set of concepts $\mathcal{A}_{\mathcal{P}_\xi}$ can, in fact, be considered as a domain model which can support later decision-making Power [2002] on the application domain. As a standard treatment is previously associated to every class by experts, evaluation of $\mathcal{A}_{\mathcal{P}_\xi}$ on new objects can help for treatment assignment. In this context, the possibility of easily interpreting and understanding the classes is critical. The proposed method provides simple and short rules which use to be easier to handle than those provided by other inductive methods.

2.1 Finding best concept at every iteration

The quality of a single rule $r : A_C(i) \xrightarrow{p} C$ is evaluated according to 3 criteria:

Support (Sup)**:** is the proportion of objects in $\mathcal{I}$ that satisfy the antecedent of the rule, Liu et al. [2000]. $Sup(r) = \frac{card\{i \in \mathcal{I} \;\; tq \;\; A_C(i)=true\}}{n}$. It measures the popularity of a rule.

Relative covering $(CovR)$**:** is the proportion of objects in class C that satisfy the antecedent of rule. $CovR(r) = \frac{card\{i \in C \;\; tq \;\; A_C(i)=true\}}{n_c}$. It measures the coverage of the rule inside a certain class.

Confidence $(p(r))$**:** proportion of objects in the antecedent $(A_C(i) = true)$ that belong to C, Liu et al. [2000]. $p(r) = \frac{card\{i \in C \;\; tq \;\; A_C(i)=true\}}{card\{A_C(i)=true\}}$. It measures the correctness of r.

The quality of a Knowledge Base is evaluated according to 3 summarizing criteria:

Average confidence: $\overline{p}(\mathcal{R}) = \frac{\sum_{\forall r \in \mathcal{R}(\mathcal{P}_\xi)} p(r)}{n_\mathcal{R}} = \frac{\sum_{\forall r \in \mathcal{R}(\mathcal{P}_\xi)} \frac{card\{i \in C \;\; tq \;\; A_C(i)=true\}}{card\{A_C(i)=true\}}}{n_\mathcal{R}}$

Total Support: $Sup_T(\mathcal{R}) = \sum_{\forall r \in \mathcal{R}} Sup(r) = \sum_{\forall r \in \mathcal{R}} \frac{card\{i \in \mathcal{I} \;\; tq \;\; A_C(i)=true\}}{n}$

Global covering: $CovG_{lobal}(\mathcal{R}) = \frac{\sum_{\forall C \in \mathcal{P}_\xi} card\{i \in C \;\; tq \;\; A_C(i)=true\} \times n_c}{n}$

Five different methods of selecting best concepts and combining with the knowledge of previous iteration are considered:

Best Global concept and Close-World Assumption (BG &CWA): Restrict the search to the set of certain rules (p(r)=1) $\mathcal{S}(\mathcal{P}^*_{\xi+1}) \subset \mathcal{S}(\mathcal{R}^*_{\xi+1})$. Choose the rule that maximizes the relative covering in $\mathcal{S}(\mathcal{P}^*_{\xi+1})$. Use a Closed-World Assumption (CWA) to conceptualize the complementary class by means of the negation of selected concept.

Best local concept and no Close-World Assumption (BL &noCWA): Choose the rule that maximizes the relative covering inside $\mathcal{S}_{C_i}(\mathcal{P}^*_{\xi+1}) = \{r \in \mathcal{S}(\mathcal{P}^*_{\xi+1}) \mid r : A_C(i) \xrightarrow{p} C_i\}$ and $\mathcal{S}_{C_j}(\mathcal{P}^*_{\xi+1}) = \{r \in \mathcal{S}(\mathcal{P}^*_{\xi+1}) \mid r : A_C(i) \xrightarrow{p} C_j\}$.

Best local concept and Close-World Assumption (BL &CWA): Choose the rule that maximizes the relative covering in both $\mathcal{S}_{C_i}(\mathcal{P}^*_{\xi+1})$ and $\mathcal{S}_{C_j}(\mathcal{P}^*_{\xi+1})$. Use a CWA to add the negation of the concept selected for the complementary class.

Best local concept and partial Close-World Assumption (BL &partial-CWA): Includes the same concepts as the Best local concept and Close-World Assumption except when the selected concept refers to the same variable for the two classes. In this case the original concept is kept.

Best local-global concept and Close-World Assumption (BL+G &CWA): Includes the same variables as the BL &partial-CWA except when the selected concepts refers the same variable for both classes. In this case the best concept is kept and the negation is added to the complementary class.

3 CASE STUDY

A case study in this paper was the pilot plant, located in Domđale-Kamnik waste water 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 m^3 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 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 waste water 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 waste water 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. The database used in this study consists of 365 daily averaged observations 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 are:

- NH4-influent: ammonia concentration at the influent of the pilot plant(pp) (3 in Fig. 2).
- Q-influent: waste water influent flow rate of the pp (7 in Fig. 2).
- TN-influent: concentration of the total nitrogen at the influent of the pp (4 in Fig. 2).
- TOC-influent: total organic carbon concentration at the influent of the pp (5 in Fig. 2).
- Nitritox-influent: measurement of the inhibition at the influent of the pp (6 in Fig. 2).
- h-waste water: height of the waste water in the tank (no Fig. 2).
- O2-1aerobic: dissolved oxygen concentration in the 1st aerobic tank (3rd tank) (12-Fig. 2).
- Valve-air: openness of the air valve (0-100%), highly related with Q-air (V2 in Fig. 2).
- Q-air: total air flow that is dosed in both aerobic tanks (1 in Fig. 2).
- NH4-2aerobic: ammonia concentration in the second aerobic tank (9 in 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 Fig. 2).
- Temp-waste water: temperature of the waste water (14 in Fig. 2).
- TOC-effluent: total organic carbon concentration at the effluent of the pp (no in Fig. 2).
- Freq-rec: frequency of the internal recycle flow rate meter (no in Fig. 2).
- FR1-DOTOK-20s (Hz): frequency of the motor that pumps the waste water into the plant.

The data base was clustered in a previous work in order to identify typical situations that could improve decision making, since managing WWTP is difficult in general and requires great expertise. See Metcalf and Eddy [2003] for details on the problematics related with management and control of WWTP and the difficulties of finding global mechanistic models. In Gibert [1996] clustering based on rules was used with the following Knowledge Base:

$$KB = \{r_1 : ((and(>= (NH4 - 2aerobic)10.0)(> (TN - effluent)18.0)) \rightarrow Mmonia),$$
$$r_2 : ((and(< \quad (NH4 - 2aerobic)10.0)(> (TN - effluent)18.0)) \rightarrow Nitrogen)\}$$

with 38 objects satisfying r_1, 80 objects satisfying r_2 and the final dendrogram of Figure 1 (see details in Pérez-Bonilla et al. [2008]). A final partition in 4 classes was $\mathcal{P}_4 = \{Cr_{353}, Cr_{357}, Cr_{358}, Cr_{360}\}$ is obtained. 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 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, as oxygen concentration in the aerobic tank is high enough, quite low effluent nitrogen concentrations can be obtained. In front of low influent flow rate, the effluent concentrations can be low if the oxygen concentration in the aerobic tanks is high.
- Cr_{358}, explains the situation when the waste water temperature is low. In this case nitrogen removal efficiency of the plant is rather low. This happens because microorganisms in the tanks do not work so intensively in cold conditions and therefore higher concentrations at the effluent can be expected.
- Cr_{360}, shows the situation when the waste water 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.

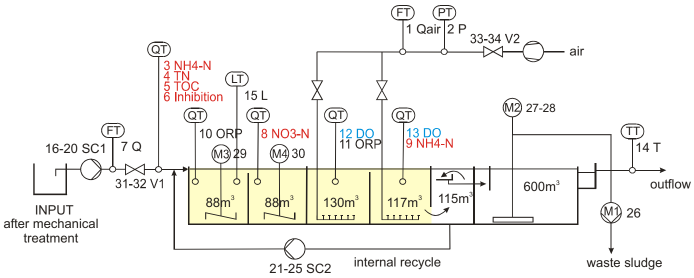

Figure 2: MBBR (Moving Bed Biofilm Reactor) pilot plant with sensors and actuators.

4 Results

In this section CCEC has been applied to the data of the plant by testing the 5 aggregation criteria presented before and compared with the interpretation provided by the experts from scratch. The descriptions obtained for every class with the different methods are shown in Table 3. Results for intermediate iterations are presented in Pérez-Bonilla et al. [2008]. Table 1 shows quality indicators of the results (Confidence, Support and Coverage). The method that gives the most similar interpretation to those provided by the expert is the *Best Local-Global and Close World Assumption (BL+G &CWA)*, which from a technical point of view also seems to represent the more equilibrated option with the second higher values in both global coverage and support. The greatest Global coverage is from *Best Local and Close World Assumption*, but this interpretation is redundant. So the best interpretation is the one obtained using *BL+G &CWA*.

5 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* 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

Best global concept and Close-World assumption:

$$
\begin{array}{lll}
\mathcal{R}(\mathcal{P}_4)=\{ & r_{Cr353}: & x_{TN-influent,i}\in[28.792,83.792]\ \wedge \\
& & x_{Q-influent,i}\in(55.666,85.092] \xrightarrow{0.45} Cr353, \\
& r_{Cr357}: & x_{TN-influent,i}\in[28.792,83.792]\ \wedge \\
& & x_{Q-influent,i}\in[49.706,55.666] \xrightarrow{0.47} Cr357, \\
& r_{Cr358}: & x_{TN-influent,i}\in[0.0,28.792)\ \wedge \\
& & x_{Temp-ww,i}\in[8.472,13.327) \xrightarrow{1.0} Cr358, \\
& r_{Cr360}: & x_{TN-influent,i}\in[0.0,28.792)\ \wedge \\
& & x_{Temp-ww,i}\in[13.327,21.896] \xrightarrow{0.86} Cr360\}
\end{array}
$$

Best local concept and no Close-World assumption:

$$
\begin{array}{lll}
\mathcal{R}(\mathcal{P}_4)=\{ & r_{Cr353}: & x_{Valve-air,i}\in(54.777,69.898]\ \wedge \\
& & x_{Q-influent,i}\in(55.666,85.092] \xrightarrow{1.0} Cr353, \\
& r_{Cr357}: & x_{Valve-air,i}\in(54.777,69.898]\ \wedge \\
& & x_{FR1-DOTOK,i}\in[42.276,44.167) \xrightarrow{1.0} Cr357, \\
& r_{Cr358}: & x_{TN-influent,i}\in[0.0,28.792)\ \wedge \\
& & x_{Temp-ww,i}\in[8.472,13.327) \xrightarrow{1.0} Cr358, \\
& r_{Cr360}: & x_{TN-influent,i}\in[0.0,28.792)\ \wedge \\
& & x_{Temp-ww,i}\in(20.928,21.896] \xrightarrow{1.0} Cr360\}
\end{array}
$$

Best local concept and Close-World Assumption:

$$
\begin{array}{lll}
\mathcal{R}(\mathcal{P}_4)=\{ & r_{Cr353}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792]")\ \wedge \\
& & (x_{Q-influent,i}\in(55.666,85.092]\ \vee \\
& & x_{FR1-DOTOK,i}\in[44.167,50.7]) \xrightarrow{0.4} Cr353, \\
& r_{Cr357}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792])\ \wedge \\
& & (x_{Q-influent,i}\in[49.706,55.666]\ \vee \\
& & x_{FR1-DOTOK,i}\in[42.276,44.167) \xrightarrow{0.48} Cr357 \\
& r_{Cr358}: & (x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & ("x_{Temp-ww,i}\in[8.472,13.327)"\ \vee \\
& & "x_{Temp-ww,i}\in[8.472,20.928]") \xrightarrow{0.3} Cr358, \\
& r_{Cr360}: & ("x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & ("x_{Temp-ww,i}\in(20.928,21.896]"\ \vee \\
& & "x_{Temp-ww,i}\in[13.327,21.896]") \xrightarrow{0.45} Cr360
\end{array}
$$

Best local concept and partial Close-World Assumption:

$$
\begin{array}{lll}
\mathcal{R}(\mathcal{P}_4)=\{ & r_{Cr353}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792]")\ \wedge \\
& & (x_{Q-influent,i}\in(55.666,85.092]\ \vee \\
& & x_{FR1-DOTOK,i}\in[44.167,50.7]) \xrightarrow{0.4} Cr353, \\
& r_{Cr357}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792])\ \wedge \\
& & (x_{Q-influent,i}\in[49.706,55.666]\ \vee \\
& & x_{FR1-DOTOK,i}\in[42.276,44.167) \xrightarrow{0.48} Cr357 \\
& r_{Cr358}: & (x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & "x_{Temp-ww,i}\in[8.472,13.327)" \xrightarrow{0.54} Cr358, \\
& r_{Cr360}: & ("x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & "x_{Temp-ww,i}\in(20.928,21.896]" \xrightarrow{0.83} Cr360
\end{array}
$$

Best Local-Global concept and Close-World Assumption:

$$
\begin{array}{lll}
\mathcal{R}(\mathcal{P}_4)=\{ & r_{Cr353}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792]")\ \wedge \\
& & (x_{Q-influent,i}\in(55.666,85.092]\ \vee \\
& & x_{FR1-DOTOK,i}\in[44.167,50.7]) \xrightarrow{0.4} Cr353, \\
& r_{Cr357}: & (x_{Valve-air,i}\in(54.777,69.898]\ \vee \\
& & x_{TN-influent,i}\in[28.792,83.792])\ \wedge \\
& & (x_{Q-influent,i}\in[49.706,55.666]\ \vee \\
& & x_{FR1-DOTOK,i}\in[42.276,44.167) \xrightarrow{0.48} Cr357 \\
& r_{Cr358}: & (x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & "x_{Temp-ww,i}\in[8.472,13.327)" \xrightarrow{1.0} Cr358, \\
& r_{Cr360}: & ("x_{TN-influent,i}\in[0.0,28.792)"\vee \\
& & "x_{Valve-air,i}\in[28.604,54.777]")\ \wedge \\
& & "x_{Temp-ww,i}\in[13.327,21.896]" \xrightarrow{1.0} Cr360
\end{array}
$$

Figure 3: Knowledge base induces by 5 methods

Table 1: Comparison among the 5 proposals

Met.	Ruler	$Concec.$	$\#\{i \in A^{\xi}_{C}\}$	$\#\{A^{\xi}_{C} \cap i \in C\}$	n_c	*p(r)*	*Sup(r)*	*CovR(r)*
BG	r_{Cr353}	$Cr353$	220	99	122	45,00%	60,27%	81,15%
	r_{Cr357}	$Cr357$	98	46	50	46,94%	26,85%	92,00%
&	r_{Cr358}	$Cr358$	6	6	93	100%	1,64%	6,45%
CWA	r_{Cr360}	$Cr360$	38	33	100	86,84%	10,41%	33,00%
	$(\overline{p}(\mathcal{R}))$					69,70%		
	$Suma$		362	184	365		99,18%	
	$(CovG_{lobal}(\mathcal{R}))$							50,4%
BL	r_{Cr353}	$Cr353$	27	27	122	100%	7,40%	22,13%
&	r_{Cr357}	$Cr357$	1	1	50	100%	0,27%	2,00%
no	r_{Cr358}	$Cr358$	6	6	93	100%	1,64%	6,45%
CWA	r_{Cr360}	$Cr360$	3	3	100	100%	0,82%	3,00%
	$(\overline{p}(\mathcal{R}))$					100%		
	$Suma$		37	37	365		10,14%	
	$(CovG_{lobal}(\mathcal{R}))$							10,1%
BL	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
&	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
CWA	r_{Cr358}	$Cr358$	299	91	93	30,43%	81,92%	97,85%
	r_{Cr360}	$Cr360$	220	99	100	45,00%	60,27%	99,00%
	$(\overline{p}(\mathcal{R}))$					40,61%		
	$Suma$		929*	360	365		254,52%**	
	$(CovG_{lobal}(\mathcal{R}))$							98,6%
BL	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
&	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
parc.	r_{Cr358}	$Cr358$	114	62	93	54,39%	31,23%	66,67%
CWA	r_{Cr360}	$Cr360$	35	29	100	82,86%	9,59%	29,00%
	$(\overline{p}(\mathcal{R}))$					56,06%		
	$Suma$		559*	261	365		153,15%**	
	$(CovG_{lobal}(\mathcal{R}))$							71,5%
BL	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
+G	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
&	r_{Cr358}	$Cr358$	114	62	93	54,39%	31,23%	66,67%
CWA	r_{Cr360}	$Cr360$	220	99	100	45,00%	60,27%	99,00%
	$(\overline{p}(\mathcal{R}))$					46,60%		
	$Suma$		744*	331	365		203,84%**	
	$(CovG_{lobal}(\mathcal{R}))$							90,7%

of *BbD* and an interactive combination of concepts upon hierarchical subdivisions of the domain. 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 and *et alt.* [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 operating the waste water treatment plant on a concrete day is obtained upon a reduced number of variables together with an estimation of the risk associated to that decision (which is related with the certainty of the rule). In this work 5 different criteria for selecting the variable to keep at every iteration are assessed and exhaustive comparison among those 5 criteria is presented, either in terms of approaching expert's descriptions as well as validating structural goodness of results, by means of confidence, support and covering: *Best Local-Global and Close World Assumption (BL+G &CWA)*, is the most equilibrated option for the technical point of view and the one with better approaches exper'ts comprehension. Comparisons of results provided by *CCEC* with *BL+G &CWA* and other methods are presented in Gibert et al. [2006]. *Logistic Regression*, *Decision Trees* and the *Discriminant Analysis* provide results much more disconnected from the interpretation proposed by the expert than the results provided by *CCEC*. Preliminary tests with other classical inductive methods evidenced the need to develop CCEC, since, as mentioned in the abstract, traditional rules-inductive methods priorize the predictive power of the final rules as well as the compacity, and usually, they hid redundant variables in the final rules. This is of course a very sound approach when prediction is the main goal of rules induction. But produces low performances for comprehensive purposes. From our experiences we could clearly observe that one of the most informative characteristics of a class, to permit expert's conceptualization, is the local relationships among redundant variables appearing in the different classes. For that reason, CCEC is a proposal that do not guarantee the most compact concepts, but provides solutions that may include class-redundant variables. Finally the methodology guarantee that important variables at the output are also included in the final description. In long term, the proposal could be extended to keep more than two variables per iteration. Formal comparison with other classical rules-inductive methods, like Prism, Rules or CN2 is currently in progress to better show these effects.

Acknowledgments We would like to thank the staff of the Domđale-Kamnik WWTP and particularly Dr. Darko Vrecko from Josef Stefan Research Institute, Ljubjiana, Slovenia. This research has been partially financed by the project TIN 2004-01368 from Spanish governemnt.

REFERENCES

Fayyad, U. and *et alt. Advances in Knowledge Discovery and Data Mining*, chapter From Data Mining to Knowledge Discovery: An overview. AAAI/MIT Press., 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. In *LNAI*, volume 1510, pages 83–92. Springer, 1998.

Gibert, K. and A. Pérez-Bonilla. Taking advantage of the hierarchical structure of a clustering for automatic generation of classification interpretations. In *4th EUSFLAT*, pages 524–529, España, septiembre 2005.

Gibert, K. and A. Pérez-Bonilla. Revised boxplot based discretization as a tool for automatic interpretation of classes from hierarchical cluster. In *IFCS 2006*, page in press, Slovenia, july 2006.

Gibert, K., A. Pérez-Bonilla, and R. G. *A Comparative Analysis of different classes-interpretation support techniques*, volume 146, pages 37–46. IOS press, 2006.

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

Hand, D. J. Classification and computers: shifting the focus. In *COMPSTAT: Proceedings in Computational Statistics*, pages 77–88. Physica-Verlag, 1996.

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

Metcalf and Eddy. *Wastewater engineering treatament. Disposal and reuse.* McGraw-Hill, 2003. 4th Ed. revised by George Tchobanoglous, Franklin L. Burton NY.US.

Pérez-Bonilla, A. and K. Gibert. Towards automatic generation of conceptual interpretation of clustering. In *LNCS*, volume 4756, pages 653–663. Springer, 2007.

Pérez-Bonilla, A., K. Gibert, and D. Vrecko. Automatic generation of conceptual descriptions of classifications in environmental domains. In *Procs.iEMSs'08*, volume 3, pages 1791–1798, 2008.

Power, D. J. *Decision support systems: concepts and resources for managers.* Westport, Conn. and Quorum Books., 2002.

Volle, M. Analyse des données, 1985. Ed. Economica, Paris, France.