Characterization and Interpretation of Classes Based on Fuzzy Rules in ill-Structured Domains

Fernando Vázquez[1], Juan Luis Díaz de León[2]

[1] UPIICSA-IPN, Computer Science Dept., México, D.F.
fvazquez_t@hotmail.com
[2] CIC-IPN, Computer Science Dept., México, D.F.
jdiaz@cic.ipn.mx

Abstract.
Nowadays, when a classification is given from a set of objects, it seems to be necessary to make use of tools to assist users to interpret tasks, in order to establish the semantic structure of the resultant classes from a given classification. It is often enough for the user to build the classes automatically, but he needs a sort of tool to help himself to understand the reason why such classes were detected there. CIADEC is a computer system that implements the methodology AUGERISD, which allows us to obtain the automatic characterization and interpretation of conceptual descriptions, combining: concepts, artificial intelligence techniques, inductive learning and statistics. A system based on fuzzy rules to find out a characterization of the given classes by an automatic form is described in this paper. A specific case applied to Wastewater Treatment Plant (WWTP) shows the stages for this methodology.

1 Introduction

The aim of this paper is to present a methodology which combines statistical tools through inductive learning, in such a way that it is the base of statistics analysis for several (numeric) measurements. Such methodology can identify the characteristic situations (classes) which can be found in the plant and it also produces a *conceptual description* of them. Once they are identified and *understood* by the user, this typical situations may be used afterwards to support the process of the decision making. This decision making may be either automatic or not.

The process that was used helps us to know that the outflow water quality (according to quality water standards) is really complex. On the other hand, due to intrinsic features of wastewater and the consequences of a bad administration of the plant, such process is complex and delicate [3].

Just to have a general idea of what is happening in the plant, we provide a very brief description of the process: the water flows sequentially through three or four stages which are commonly known as pretreatment, primary, secondary and advanced treatment (see [4] for a detailed description of the process). Figure 1 depicts its general structure.

© A. Gelbukh, R. Monroy. (Eds.)
Advances in Artificial Intelligence Theory
Research on Computing Science 16, 2005, pp. 261-270

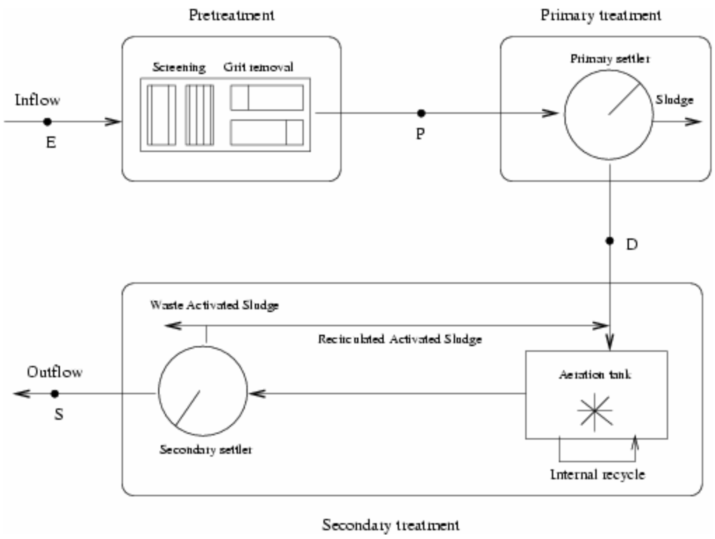

Fig. 1. General structure of wastewater treatment plant.

The paper is organized as follows: In section 2 the presentation of the Wastewater Treatment Plant (WWTP) data is shown; in section 3 the methodology for the characterization of the classes is described; in section 4 the methodology is applied to a WWTP; in section 5 the interpretation of the results obtained of the automatic characterization and interpretation of conceptual descriptions taken as references from the partition of Klass+ from which four classes are displayed. Finally, in section 6 the conclusions and future research are presented.

2 Presentation of the Wastewater Treatment Plant (WWTP) data

The set of qualitative and quantitative data analyzed in this research was taken from a treatment plant on the Catalonian coast (Lloret city, Spain). It was made up of 218 observations taken from the same number of consecutive days. Some of the observations were taken at different stages of the process, whereas others were taken based on calculations from those. The first ones correspond to the daily average of repeated measurements on a total of 63 attributes that were measured at various points in the facility (AB: the entrance of the center, SP1: the outlet of the first settling tank, B: in the biological reactor, SP3: in the third settling tank, AT: the treated water), as well as a description of the plant’s state at the moment of measurement.

3 Methodology for the characterization of the classes

The objective is to generate a fuzzy rules system from a set of data which has been described by the attributes above mentioned; and to obtain a description that can be easily understandable for the user, and which indicates particularities from each class among the rest of them in a given partition $P = \{c_1, \ldots\ldots, c_\xi\}$.

3.1 Statistical description of the attributes

In this first stage, some classical descriptive techniques are used in order to identify the behavior and nature of the data referring to the data matrix X . This stage is useful to obtain preliminary information about the variability of the measurements and to represent a multiple box-plot which will let us observe the relation between the attributes and the classes, and it is especially useful to represent the difference among clusters or classes.

3.2 The use of multiple box-plots as a graphic tool for the detection of characterizing attributes.

As we have already pointed out, the *multiple box-plots* are excellent base for this research, as a tool to view and compare the distribution of an attribute through all the classes. In this representation, it is possible to identify the characterizing attributes from class C , explained through the concept of proper value from a class C . It is quite simple to graphically observe if the multiple box-plot of a certain class doesn't intersect others; in such a case, the attribute is fully characterizing[1]. Sometimes, it is only a part of the box-plot that is not intersected; in this case it means that we have a partial characterizing variable.

3.3 Study of classes interactions

In this process, it is utterly important to consider the attributes, in their natural state, avoiding any arbitrary transformation about their nature that could change the sense of the intersection. This stage consists on identifying all the intersections occurring among the values of attributes and the different classes. We determine at what point in the range of attributes these intersections are changing, thus allowing identification of the different combinations existing among classes where the same value of a certain variable or attribute can be found and, consequently, to let proper values emerge (characterizing values). This would tell us whether they were fully or partially characterized.

3.4 Space discreteness of attributes

Exact intersections can be found with minimal computational cost, by simply calculating the minimum and maximum values by variable and class, and ordering

[1] Onwards, we will simple call characterizing variable of that class.

them. From this order, the discreteness of the variable is defined by a set of intervals of variable length $I^k = \left\{ I_1^k + I_2^k, \ldots\ldots, I_{2\xi-1}^k \right\}$ so that $U_{s=1}^{2\xi-1} I_s^k = D_k$, in which the *unique values* of a variable in all the classes can be identified.

To extend these concepts, thus, if m_C^k and M_C^k are the minimum and maximum of the variable X_k in the class $C \in P$, which have been observed in the descriptive of the multiple box-plots, where $m_C^k = \min_{i\in C}\{x_{ik}\}$ and $M_C^k \max_{i\in C}\{x_{ik}\}$. Now, we order them in ascendant form, and this is:

- Define M^k like the set of all the minimums and maximums corresponding to the variable X_k, in all P classes, this is:
  $M^k = \{m_{c1}^k, \ldots\ldots, m_{c\xi}^k, M_{c1}^k, \ldots. M_{c\xi}^k\}$, being the $Card(M^k) = 2\xi$
- Ordering M^k from minimum to maximum, a set is constructed Z^k in the following way: $Z^k = \{z_i^k ; i = 1, \ldots 2\xi\}$, so:
  i. $z_1^k = \min M^k$ and
  ii. $z_i^k = \min(M^k \backslash \{z_j^k ; j < i\}), i = 2, \ldots\ldots, 2\xi$

  Since $Z^k = \{z_i^k\}$ is an ordered set, its elements have the following property: $Z^K = \{z_J^K \mid z_{J-1}^K < z_j^k : 1 < j < 2\xi\}$, this set is called *cut points.*
- From this ordered set, we build the intervals system of variable length I^k in the following form: $I^k = \{I_s^k : 1 \le s \le 2\xi - 1\}$, in which:
  i. $I_1^k = [\, z_1^k, z_2^k \,]$
  ii. $I_s^k = (\, z_s^k, z_{s+1}^k]$, with $s = 2, \ldots, 2\xi - 1$

A new categorizing variable is then defined as I^k; whose set of values is $D^k = \{I_1^k, \ldots\ldots, I_{2\xi-1}^k\}$. I^k identifies all intersections among classes that X_k defines, representing a system of length intervals associated to such variable. Thus, were 2ξ different cut points, then $2\xi - 1$ intervals the most are generated and $Card(D^k) = 2\xi - 1$, taking into account that ξ is the number of initial classes of reference that we want to characterize.

On the other hand, since D^k is the domain of X_k, D^k represents the categorization of itself, yet not arbitrary at all, it is also calculated immediately. Finally, it is necessary to observe that in order to construct I^k it is not necessary to perform the multiple box-plots anymore, even though it is still an excellent representation of what it is being done.

3.5 Construction of distribution table of classes versus intervals

From a system of intervals, a contingency table is set for a variable X_k, as a matrix A set of numbers in which each line represents an interval I^k and each column represents a class found in the previous stage, for the classes reference partition P. Therefore, any given cell in this matrix indicates the number of elements in the domain I, whose values of X_k are found in an interval represented by I_s^k.

In general, for a given value of the variable X_k, objects from different classes are found.

3.6 Generating a fuzzy rules system $R(X_k, P)$

From this distribution from table *B*, a rule system is constructed for each non-null cell p_{sc}. From this matrix, the following rules are generated: if $X_{ik} \in I_s^k \xrightarrow{psc} C$.

This way, $R(X_k, P)$ can be used to recognize the class (or classes) belonging to a certain day in which $i = (x_{i1}, \ldots.., x_{iK}, \ldots., x_{ik})$ belongs to, according to its value of X_k.

3.7 Interpretation of resultant classes

The interpretation of the resultant classes tends to be utterly important to use the generated knowledge as an aiming tool for the future decision taking. In fact, it has been remarked that the validation of a classification has been considered as the degree of interpretability and/or the utility of these, without any other criteria but that of a specialist that looks at the resultant classes.

Having the conditioned table of distributions as a base to get the previously described intervals in 3.4 from this section, any spare part *I* can be associated to its degree of belonging to each class. This info gives us a graphic of diffused degrees of belonging for each class and for each variable shown in figure 3. In the graphic, the horizontal axe is common and represents the range X_k. For every single class the degree of belonging of values of k is represented according to the rules. The scaffolding shape for such functions of belonging must be categorized from X_k in I^k. Thus, given a value from X_k, it is easily noticeable its relation to other classes.

4 Application of the methodology to WWTP data

In this section, the methodology AUGERISD [5] is applied to the data that were gathered from the Wastewater Treatment Plant with Klass$^+$ [1] being partitioned into four classes. We applied the methodology automated through system CIADEC [6] to identify the relevant characteristics of the reference partition, having a fuzzy rules system obtained that will allow us to get the characterization and interpretation of the conceptual descriptions for the resulting classes of such reference partition, considering the analysis of all the attributes in such a way.

5 Generation of Interpretations

In [1] and [2] a first approach for an efficient algorithm which generates automatic interpretations of the classes is introduced. In this research, the main work is about the interpretation of the classes on the basis of categorical attributes. In this section, we are interested in the use of numerical attributes for interpreting classes. The real application we presented here is especially indicated for this goal, since categorical attributes are not presented in the data matrix at all and only numerical measurements are useful to describe data.

From this automated methodology we obtain the conditioned distributions to each interval (see Table 1). The table, which has been mentioned above, gives us the percentage of elements of certain interval in each class, obtaining a reduced fuzzy rules system (see figure 2).

Table 1. Conditional Distribution Table of Attribute $Q-AB$

Intervals	Classes			
	C1	C2	C3	C4
[4910, 5881]	0.94	0.00	0.06	0.00
(5881, 6277]	0.85	0.07	0.08	0.00
(6277, 13454]	0.41	0.23	0.36	0.00
(13454, 13563]	0.00	0.34	0.33	0.33
(13563, 13563]	0.00	0.00	0.00	0.00
(13563, 14375]	0.00	0.50	0.50	0.00
(14375, 23394]	0.00	1.0	0.00	0.00

Having any of them, since it is based on the conditioned distributions table to the intervals, they can be associated with an object (day) that is called: a belonging degree for each class. This idea gives place to a graph of a belonging fuzzy degree for each class and for each attribute (see figure 3).

$$\begin{aligned}
r_1 &: x_{Q-AB,i} \in [4910\ ,5881\ ] \xrightarrow{0.94} C1 \\
r_2 &: x_{Q-AB,i} \in (5881\ ,6277\ ] \xrightarrow{0.85} C1 \\
r_3 &: x_{Q-AB,i} \in (6277\ ,13454\ ] \xrightarrow{0.41} C2 \\
r_4 &: x_{Q-AB,i} \in (13454\ ,13563\ ] \xrightarrow{0.34} C2 \\
r_6 &: x_{Q-AB,i} \in (13563\ ,14375\ ] \xrightarrow{0.5} C2 \\
r_7 &: x_{Q-AB,i} \in (14375\ ,23394\ ] \xrightarrow{1.0} C2
\end{aligned}$$

Fig. 2. Fuzzy Rules Reduced System of Attribute $Q-AB$

In this way, from a system with a method of creation of linguistic labels we automatically generate conceptual descriptions for these classes.

Using this methodology, the following interpretations are produced. The opinion of the experts is also included in the discussion below, as they confirmed the understandability of the discovered classes.

Class 1:

The output water is clean. The input water is low dirty (values are low in almost all the attributes). Experts identified this class with the class of those days with very good plant performance, as a consequence of the good conditions, even ammonium is reduced.

Class 2:

High wastewater inflow. The water which comes in is medium dirty (intermediate values in almost all the attributes: suspended solid total, chemical organic matter, biodegradable organic matter, etc). Settler is making high effect (levels of suspended and volatile suspended solid are significantly reduced at the primary treatment). From the expert point of view, this class is interpreted as the class of days with general good performance, but in which punctual problems can produce some isolated parameter with high values.

Class 3:

The water that comes in is very dirty. The output water with intermediate measures all the attributes. Nonetheless, the performance of the plant is not so good. This class contains some days in which isolated control parameters overcome the permitted values.

Experts identified this class to the class of those days with organic material overloading on summer days with optimal WWTP-operation.

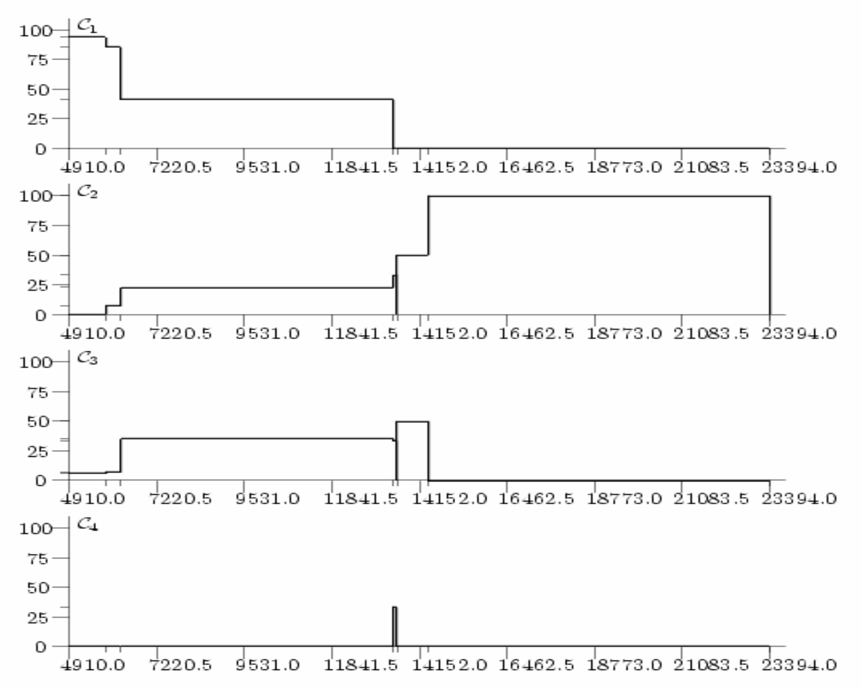

Fig. 3. Attribute Graphic $Q-AB$

Class 4:

High levels of chlorine and conductivity. From the expert point of view, this class is interpreted as the class of days with chlorine over amount.

6 Conclusions and Future Work

This work constitutes a positive experience in terms of establishing a formal methodology to automatically obtain conceptual interpretations of the classes, on the basis of numerical attributes used to describe objects (days in this case).

From a very small and partial set of rules, the system created a new level of abstraction that catches the nature of a Wastewater Treatment Plant for the given set of data (noisy, incomplete and heterogeneous), producing a set of identified classes, as well as their conceptual interpretation, which was directly interpreted by the experts on this matter.

In the future, it is our intention to explore the series of time of the behavior of the plots in their variables.

Acknowledgements
This research has been partially financed by COFAA and IPN, México.

References

1. Gibert K. In L'us de la informació simbólica en *la automatizació del tractament estadístic de dominis poc estructurats*. Ph D. Thesis, UPC, BCNA, 1994.
2. Gibert K., Aluja T. and Cortés. *Knowledge discovery with clustering based on rules, interpreting results*
3. Gimeno J.M., Béjar I., Sànchez-Marrè and Cortés U. In *"Discovering and modelling process change: An application to industrial processes"*. Practical Applications of Data Mining and Knowledge Discovery. 1997.4. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
4. Metcalf and Eddy Inc., In *Wastewater engineering: treatment/disposal/reuse*. McGraw-Hill, 1991.
5. Vázquez F., Gibert K. In *Automatic Generation of Fuzzy Rules in ill-Structured Domanis with Numerical Variables*, publisher: UPC, LSI, Report num: LSI-01-51-R. Barcelona, España. 2001. E-mail: http://www.lsi.upc.es.
6. Vázquez F., Gibert K. In Implementation of the methodology "Automatic Characterization and Interpretation of Conceptual Descriptions in ill-Structured Domains using Numerical Variables", publisher: UPC, LSI, Report num: LSI-02-28-R. Barcelona, Spain. 2002. E-mail: http://www.lsi.upc.es.
7. Rodas J., Alvarado G. and Vázquez F. Using the KDSM methodology for knowledge discovery from a labor domain. (SNPD2005). Towson University. Towson, Maryland, USA. May 2005.