Robustness of Class Prediction Depending on Reference Partition in ill-Structured Domains

Fernando Vázquez Torres[1] and Karina Gibert Oliveras[2]

[1] Technical University of Catalonia, Department of Computer Science, ⋆⋆⋆
fvazquez@lsi.upc.es
[2] Department of Statistics and Operations Research, UPC, †
karina@eio.upc.es
Barcelona 08034, Spain

Abstract. In this paper, an analysis of the robustness of classes prediction given a reference partition of a data set, applying the methodology of Automatic Generation of Fuzzy Rules in ill-Structured Domains (AUGERISD) is presented. A specific application on a WasteWater Treatment Plant (WWTP) illustrates the stages of this analysis. The methodology is based on the combination of statistical tools and inductive learning, in such a way that the nature of the data is preserved, avoiding previous transformations of the variables. Thus qualitative and quantitative information can be induced from data. This information is useful for the automatic generation of a system of fuzzy rules, which, in turn, allows the later recognition of the reference classes. In this work, we have started from a WWTP data set, two reference partitions obtained one from the classifier $LINNEO^+$ and the other from the classifier $KLASS^+$, a new proof set of objects P_0, the objective is to predict the class of these objects, the $LINNEO^+$ partition using for the time being, only numerical variables. Although it is a classification with a valid interpretation from experts, and it has been obtained an automatic classification, the experts recognize that it is not the only one possible solution and that it could even be improved there. That is why in this work, second partition produced by $KLASS^+$ is introduced. Comparison between them allows to analyze the robustness of class prediction depending on reference partition in ill-structured domains.
Keywords: Machine Learning, Knowledge Discovery, Knowledge-based System and Decision Support Systems.

1 Introduction

The objective of this paper is to study the robustness of class prediction on reference partition, starting from a partition of an ill-structured domain, applying the methodology "Automatic Generation of Fuzzy Rules in Ill-Structured Domains (ISD) with Numerical Variables" (AUGERISD) [14], that allows to characterize the different classes and to use this characterization for future predictions.

⋆⋆⋆ Scholarship holder of IPN and SUPERA, México.

† This research was partially financed by the TIC2000-1011 project: Development of an intelligent system for knowledge management of environmental data bases.

Once we assume that the reference partition is sound, it identifies typical situations in a given process and it allows to establish decision criteria in real time to manage the process itself; as well as to obtain excellent information which is later useful in the decision making which is very interesting in the context of ISD, since they are very complex and decisions have to be made in conditions of uncertainty, imprecision, noise, etc. But, very frequently, validity of the reference partition is also subjected to uncertainly, and cannot be armed. What we expect in these cases is to be able to obtain, at least, robust predicted classes for new objects. In this paper, we want to analyze among other questions, how sensible the final predicted classes are depending on the reference partition in a data matrix X.

1.1 Methodology for the class prediction

Let us call $\mathcal{I} = \{i_1, \ldots, i_n\}$ the set of individuals to be analyzed, described by a set of attributes or variables X_k. The domain or set of possible values of the numerical variables or $D_k \subseteq R$. The values of these variables X_k taken by each of the individuals $i \in \mathcal{I}$ are represented by means of rectangular matrix $\mathcal{X}$ of dimensions (n, K), of the form: $\mathcal{X} = [\ x_{ik}\ \ (i \in \mathcal{I}), (k = 1 : K)]$, where x_{ik} is the value that the *ith* individual takes for the *kth* variable. The methodology [14] presented in this research start with data matrix X and a partition of $\mathcal{I}$, $\mathcal{P} = C1, \ldots, C_\xi$ where $\forall\ \ C \in \mathcal{P},\ \ C \subseteq \mathcal{I}$. The objective is to generate a set of rules able to recognize this partition by using the given variables in such a way that the more likely classes of a new object can be stated.

The methodology consists of the following steps:

1. **The use of the multiple box-plot as a graphic tool for the detection of characterizing variables**
   The multiple box-plot [13] is an excellent starting point for this work (in the next section, it will be presented in detail). In this work it is used to identify what will be called *characterizing variables* of class C, based on the concept of the *proper value* of a class. It is quite simple to graphically observe in a box-plot if the values of variables a certain class do not intersect with other classes; in this case the variable is *totally characterizing*[3] the class. Sometimes, it is only a part of the box-plot that is not intersecting with other classes, in this case a *partially characterizing* variable is found.
   In order to identify these variables, exclusive values that the X_k variable take for class C are looked for. To do so, class interactions have to be analyzed.
2. **Discretization of attribute space**
   Exact intersections among classes can be found with minimal computational cost, simply by calculating the minimum and maximum values by variable and class, and globally ordering them. From this ordering, a discretization of the variable X_k is induced, and numerical variables X_k are converted into a set of intervals of variable length $\mathcal{I}^k = \{I_1^k, I_2^k, \ldots, I_{2\xi-1}^k\}$, so that

[3] Characterizing variable of that class for short.

$\cup_{s=1}^{2\xi-1} I_s^k = \mathcal{D}_k$. Using $\mathcal{I}^k$, the *proper values* of a variable in all classes can be identified.

3. **Construction of conditional distribution table**
   From a system of intervals $\mathcal{I}^k$, a table is set up for a variable X_k, as a matrix A in which each row represents an interval I_s^k and each column represents a class C of the reference partition P. Therefore, any given cell in this matrix indicates the number of elements in C 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 various classes are found.
   Now it is relatively easy to construct the matrix B, the matrix of distributions of C conditioned to each interval I_s^k, so that the cells represent the percentage of objects in C with respect to the element with values of X_k in I_s^k: $p_{sc} = n_{sc}/n_{I_s^k}$, where n_{sc} is the number of individuals that belong to I_s^k and class C, and $n_{I_s^k} = \Sigma_{c=1}^{\xi} n_{sc}$ is the total number of objects that are in the same interval I_s^k.
4. **Generating a system of fuzzy rules $\mathcal{R}(X_k, P)$**
   From this distribution table B, a rule system is induced for each non-null cell p_{sc}. From every cell, the following rules are generated: If $x_{ik} \in I_s^k \xrightarrow{p_{sc}} C$.
   This way, $\mathcal{R}(X_k, P)$ can be used to recognize the class (or classes) that a certain day, $i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{iK})$ is likely to belong to, according to it is value of X_k:
   - Assign x_{ik} to the corresponding I_s^k. Repeat for every variable X_k.
   - Choose the rules with I_s^k antecedents. The class(es) assigned to the object i will be those that are on the right–hand sides of those rules with probabilities p_{sc}.
5. **Validation of the total system of rules**
   In the proposed methodology, the box-plots for the determination of the characterizing values have been used, on the basis of an interval system of variable length. That allows the identification of the natural structure that underlies in the data base of the WasteWater plant variable by variable. In fact, it has allowed us to develop a fast method to construct a system of fuzzy rules associated to each variable, which can be found in the table of conditioned distributions $B = \mathcal{P}|I^k$. It is a first proposal, still in phase of development, which its developmental stage has reduced the inherent ambiguity in the system of fuzzy rules $\Re(X_k, \mathcal{P})$ on the basis of the criterion of maximum probability, which gives rise to reduced system $\Re'(X_k, \mathcal{P})$ much smaller in number of rules, and is non ambiguous although it does contain a level of uncertainty.

The paper is organized as follows: In §2 the application of a WasteWater treatment plant is presented; in §3 the presentation of reference partitions is described; in §4 the classes prediction with the system CIADEC (Automatic Characterization and Interpretation of Conceptual Description) is made; in §5 the results obtained of the assignment of classes to the objects for the set test taking as references the partitions of $Linneo^+$ and $Klass^+$ are displayed; in

§6 the comparison inductive method to extract knowledge in WWTP data is discussed; in §7 the conclusions are discussed. Finally, in §8, future research is presented.

2 Application to the WasteWater Treatment Plant data

To adequately deal with WasteWater, various operations and unitary processes are necessary. These processes include various combinations of physical, chemical and biological agents. Figure 1 represents a typical diagram, as well as the logical treatment sequence, divided into different phases. The following is a brief summary (for more detail, refer to [12]) of the process.

The first stage of purification is the **pretreatment** of the WasteWater. In this phase, the solid waste is dredged and separated when it reaches the collector. This is done to avoid posterior obstructions and other trouble with the pumps and valves that are used throughout the process.

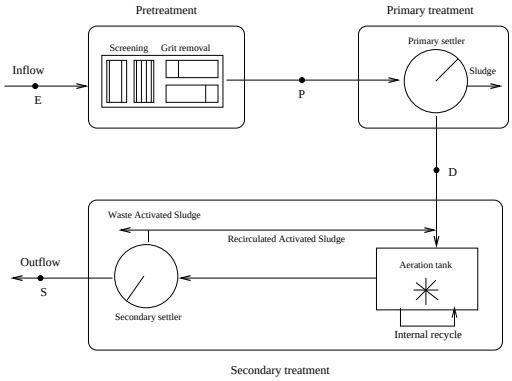

Fig. 1. Typical diagram of WasteWater treatment process

The **primary treatment** corresponds to the second stage of the process. Here, the water is left in a primary settler in order to gather as much of the organic material as possible that got through the pretreatment: the rest of the sand and inorganic particles.

The **secondary treatment**, the most important stage of the process is carried out next. Here, the biological process of the water is accelerated; i.e. the breaking down of the dissolved organic material in the WasteWater. This occurs due to a multispecific population of microorganisms, know as *biomass*, acting on it.

Finally another decanting process is carried out on the secondary sediment, to be later released into the water. The goal of the aforementioned process is to achieve an adequate separation between the already-treated water and the biomass that becomes processed with the sludge and refeeds the biological reactor.

Therefore, it becomes vitally important to implement an automated system that provides relevant information about the situation in the plant at any given moment. Our research lies in contributions along these lines.

2.1 Presentation of the data

The set of qualitative and quantitative data analyzed in this work was taken from a WasteWater Treatment Plant on the Catalonian coast (Spain). It is made up of 218 observations taken from the same number of consecutive days. Each observation contains measures in six different points of the plant, as well as others based on calculations from those first ones. Data correspond to the daily average of repeated measurements on a total of 63 variables which describe the state of the plant at that moment; some measures were repeated in various points in the plant (AB: the entrance of the plant, SP1: after the first settler, B: in the biological reactor, SP3: after the third settler, AT: the treated water, E: the emitted water).

The data set $\mathcal{I}$ of 218 days, described by the 17 numerical variables recommended by the expert and previously classified was considered as a training set T_0 and other a data set of 25 individuals also previously classified will be used as for evaluation a test set P_0.

3 WWTP data classifications: $Linneo^+$ $\mathcal{P}_L$ and $Klass^+$ $\mathcal{P}_K$

The state of the plant (the class attribute) was previously identified by means of a semi-automatic classification process using the $Linneo^+$ [1] tool and expert criteria.

$Linneo^+$, which is a semi-automated knowledge acquisition tool concerned with building classifications for ill-structured domains, was the software used for clustering the data.

$Linneo^+$ is an unsupervised learning (clustering) method that determines useful subsets or classes of the data. In the classification step, $Linneo^+$ works by defining a space of n dimensions, where n is the number of variables included in the database (in this case 63).

After an iterative classification process supervised by the expert, the 218 days were classified into 20 situations occurring in the plant. These 20 classes correspond to the clusters obtained with the same $Linneo^+$ classification using a radius equal to 10 except for two classes undetected with this radius. Other classifications with different radius discovered the other two new clusters corresponding to states of the plant.

Although the classification of $Linneo^+$ is a classification that the experts have given a valid interpretations to, it has also been observed as an automatic classification process, the experts themselves recognize that it is not the only one and could be improved. The validation of this classification has not even been

contrasted by any objective means, which is reason why in this work a second reference partition obtained by the $Klass^+$ is proposed.

$Klass^+$ [3] is a clustering tool that was originally designed to make comparisons between classic statistical methods and a system oriented to the fuzzy knowledge based classification and becoming an independent tool (*Linneo* [10]) that allows the classification of ill-structured domains.

The $Klass^+$ system presents important differences to other classifiers: the processing of symbolic information and a specific classification methodology of with declarative restrictions, thus, $Klass^+$ is situated more like a tool to aid knowledge acquisition, with a dual purpose:

- a classification method with knowledge-based restrictions is implemented.
- a tool to aid the acquisition based on statistical methods is constituted that provides the automatic generation of rules for a oriented system to diagnosis and/or prediction.

The methodology of $Klass^+$ is based on a method of hierarchic ascending classification, using the algorithm of chained reciprocal neighbours . This strategy of classification consists of detecting the pairs of reciprocal neighbours that can be fused and thus constructing the aggregations tree. Additionally, it works with the following metrics: Euclidean, standardized Euclidean, χ^2 (chi-square), mixed metrics [3], Gower [4], Ralambondrainy [11], Diday-Gowda [2] and Generalized Minkowski metrics [9].

The $Klass^+$ classification for the 218 objects of the training set was made taking into account the rules given by the expert and making a 20-class cut on the general hierarchical tree.

3.1 Comparing the $Linneo^+$ and $Klass^+$ classifications

Of the classification obtained by the $Klass^+$ tool, after analyzing the elements that contain each one the 20 classes, a comparison with the partition $Linneo^+$ is established. The obtained results are:

- Of the 20 classes between both classifications, seven classes are identified as very similar classes, the relations of these classes between the partitions of $Linneo^+$ and $Klass^+$ are:$C02$ with $\hat{C}05$, $C09$ with $\hat{C}09$, $C10$ with $\hat{C}10$, $C11$ with $\hat{C}11$, $C12$ with $\hat{C}12$, $C14$ with $\hat{C}14$ y $C17$ with $\hat{C}17$, which represents a relative coinciding of its elements of 86.36%;
- A matrix is obtained making a cross table between the two classifications where the elements of the diagonal represent the common elements between the different classes from both classifications having a total of 94 objects in similar classes that represent 43.11% of coinciding elements. The rest of the objects are dispersed in the other classes, forming different classes with different characteristics.

4 The class prediction with CIADEC

CIADEC (Characterization and Automatic Interpretation of Conceptual Descriptions in ill-Structured Domains using Numerical Variables) is a system that implements the AUGERISD [14] methodology "Automatic Generation of Fuzzy Rules in ill-Structured Domains with Numerical Variables", which allow to obtain the automatic characterization and interpretation of conceptual descriptions in ill-structured domains partitioned previously, combining concepts, techniques of artificial intelligence and statistics.

In addition, the automatization of this methodology offers a set of tools that allows us:

- to construct a rules system for the classes prediction, situations diagnosis, etc.
- to visualize the membership functions of a variable X_k to the different classes.
- to evaluate a new set of objects according to the generated rules.
- to validate the quality of the class assignment having a test set P_0.

In this research CIADEC has been used for the class prediction of new object test set having like reference the partitions that $Linneo^+$ and $Klass^+$ have obtained.

4.1 Application of the methodology to WWTP data

In this subsection, the methodology AUGERISD is applied to the data gathered from the WasteWater treatment plant with $Klass^+$ partition.

A descriptive statistic where preliminary information was obtained on: the minimum and maximum number of objects by class, ends by class and variable, the average, the median, the outliers observations, the variability of the measurements, the behavior of a variable through the processing; immediately, the knowledge to prior of the expert was considered, using a classification based on rules to take into east account knowledge and to obtain a set of classes induced by these rules. Next we applied the methodology automated through system CIADEC to identify the relevant characteristics of the resulting classes of the reference partition, being obtained a system of rules that allow us to obtain the characterization and interpretation of the conceptual descriptions of the resulting classes of the reference partition. Considering the analysis of all the variables in joint form.

5 Results

Thus, the validation process of total rules system consists: from a test set P_0 previously classified, to measure the quality of class assignment (percent of objects well-classified) to the new objects, considering an analysis in which the information aggregation of the variables in joint form is taken. This, in order to predict the membership class of each new object and to estimate the quality of this prediction, considering the aggregations criteria: maximum probability

(MP), voting (Vot) and maximum sum (MS). Once determined the class prediction of Po, these are compared with the reference classes and the number of objects well-classified calculates to obtain the "prediction accuracy" on test set Po. The obtained results are in the Table 1.

Table 1. Assignment of prediction classes considering the aggregations criteria of maximum probability (MP), voting (Vot) and maximum sum (MS) and $Klass^+$ classification (K).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
K	1	7	7	10	1	1	1	7	2	2	2	11	2	2	1	2	11	1	1	1	1	1	1	1	2
MP	15	7	11	10	1	1	1	5	1	13	1	11	2	19	2	2	2	1	1	1	1	1	1	5	15
Vot	4	7	5	10	1	1	1	2	2	2	2	11	2	2	10	2	1	2	1	1	2	1	1	2	7
MS	4	7	2	10	1	1	2	2	2	2	2	11	1	2	10	2	11	2	10	1	13	1	1	2	2

And the prediction accuracy on the test set for each one of the criteria applied to the WWTP with the $Klass^+$ reference partition, these are: 45% for MP, 36% for Vot and 40% for MS. For the $Linneo^+$ reference partition with such data, the prediction accuracy are: 48% for MP, 38% for Vot and 40% for MS. Best accuracy prediction was obtained with Voting aggregation criterion for this data base.

6 Comparison of the inductive methods for knowledge discovery in WWTP

Thus, in the article "Knowledge Discovery by means of inductive methods in WasteWater treatment plant data "'whose authors are Commas et al., published in AI Communications 14 (2001) 45-62 and where the comparison of inductive methods for knowledge discovery in WWTP were discussed, these were: C4.5 (decision trees), two methods of rule induction , the CN2 and the methodology based on boxplot, AUGERISD; two inductive methods memory-based learning (instances IBL and the cases CBL) with respect to the following parameters: number of attributes, number of examples, criteria of exactitude of classification and prediction and meaningful interpretation of the of the classes on the part of the experts. In that comparison this methodology AUGERISD performs worse that the other; with the improvements to this methodology (implementation in automatic form, combination of variables and aggregation criteria) again it returns to validate itself (use cross validation, ten folds) obtaining better results, locating to this method with best performance, as far as exactitude criteria talks about at least. Table 2 shows these last results.

Table 2. Comparison inductive methods for prediction and meaningful interpretation in WWTP data. March, 2002

Method	Number of attributes	Number of examples	Prediction accuracy on test set(%)	Meaningful interpretation	Prediction accuracy on whole data set(%)
C4.5	24	243	63.51	Partially	89.7
CN2	44	243	63.98	Partially	98.8
AUGERISD	63	243	64.5	Mostly	97.24
k-NN	63	243	76.38	No	100
J48	–	243	64.4	Partially	–
J48, bagging 10 i	–	243	70.7	No	–
J48, AdaBoost 10 i	–	243	73.6	No	–
C4.5	11	243	65.11	Mostly	87.2
CN2	19	243	65.45	Mostly	95.9
k-NN	19	243	71.22	No	100

7 Conclusions

- The methodology's computational cost is low in relation to the information it provides. This is due to the fact that it resolves an analysis of intersections between classes of degree ξ (in our case 20), calculating ξ maximums and ξ minimums and sorting them. With this method, all the points where the intersection is changed among elements in $\mathcal{P}$ are found. In addition, all the possible intersections are obtained.
- The obtaining of a boxplot-based rule induction method, that can be used in a process of supervised learning.
- Robustness of class prediction depending on reference partition in ill-structured domains.
- The application of this first methodological approach to a WasteWater Treatment Plant has allowed us to gain further knowledge concerning the various situations that present themselves in the class characterizing process.
- This work constitutes an important point in the construction of diagnosis system in WasteWater Treatment Plants. A method was presented that generates a system of fuzzy rules from numeric variables measured in the plant with the goal of identifying standard situations.
- The benefits of the methodology are:
  - To aid in the interpretation of the plant situation in any given day.
  - The knowledge gained in plant behavior by characterizing the various specific situations that can arise in the plant.
  - To provide help in the management of the plant.

8 Future Work

- To implement a fuzzy model that allows the creation of linguistics labels to automatic generating of conceptual descriptions.
- Study and selection of an aggregation criterion that allows to increase the efficiency of rule system.
- A characterization and interpretation automatic system of the conceptual descriptions of the resulting classes of a reference partition.
- As a result, the overall coverage of the system would increase.
- Application of the methodology in other ill-structured domain.

References

1. Béjar Javier. Knowledge acquisition in ill-structured domains, PhD Thesis, Depto. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. 1995.
2. Diday E. and Gowda K.C. Symbolic clustering using a new similarity measure. In IEEE Trans. on systems, man., and cib., volume 22, pages 368-378, 1992.
3. Gibert K. L'ús de la informació simbòlica en la automatizaci'o del tractament estadístic de dominis poc estructurats. Ph D. Thesis, UPC, BCNA, 1994.
4. Gibert K., U. Cortés. Clustering based on rules and knowledge discovery in ill-structured domains, Computación y Sistemas, México, 1998.
5. Gibert K. A computational technique for comparing classifications and its relationship with Knowledge Discovery. Research in official Statistics Journal. Octubre 1998.
6. Gibert K., Roda R., Cortés U., Sànchez-Marrè M. Identifying characteristic situations in wastewater treatment plants, in: Workshop in Binding Environmental Sciences and Artificial Intelligence, eds., ECAI, Berlin, 2000, pp 1-9.
7. Gibert K., Salvador A. Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. pags. 497-502, Congreso Español Sobre Tecnologías y Lógica Fuzzy, ESTYLF 2000.
8. Gower J.C. A general coefficient if similarity and some of its porpieties. Biometrics, 27:857-874, 1971.
9. Ichino M. and Yaguchi H. Generalized Minkowski Metrics for Mixed feature-type data analysis. IEEE Transaction on systems, man and cybernetics, 22(2):146-153, 1994.
10. Martín, M., LINNEO eina per a lájut en la construcció de bases de coneixement en dominis poc estructurats. Tesi llicenciatura. Departament dĹSI, UPC, 1991.
11. Ralambondrainy H.A. A conceptual version of k-means algorithm. Pattern Recognition Letters., 16:1147-1157, 1995.
12. Sànchez-Marrè, M.: DAI-DEPUR: An Integrated Supervisory Multi-level Architecture for Wastewater Treatment Plants. PhD thesis, Departamento de Lenguajes y Sistemas Informáticos de la UPC. Barcelona, Hivern de 1995/1996.
13. Tuckey J.W. Exploratory Data Analysis. Addison-Weseley, 1977.
14. Vázquez F., Gibert K. Automatic Generation of Fuzzy Rules in ill-Structured Domanis with Numerical Variables, publisher: UPC, LSI, Report num: LSI-01-51-R. Barcelona, España. 2001.
15. Vázquez F., Gibert K. Generación Automática de Reglas Difusas en Dominios poco Estructurados con Variables Numéricas. Asociación Espaola para la Inteligencia Artificial, CAEPIA-TTIA 2001. Gijón, España.