A Tipicity Concept for Data Analysis and Its Application to Cleft Lip and Palate

Leticia Vega-Alvarado[1] and Martha R. Ortíz-Posadas[2]

[1] Centro de Ciencias Aplicadas y Desarrollo Tecnológico,
Universidad Nacional Autónoma de México, Circuito exterior s/n,
Cd. Universitaria, 04510, Coyoacán, D.F. México
leticia.vega@ccadet.unam.mx
[2] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana,
Iztapalapa, México
posa@xanum.uam.mx

Abstract. The paper presents a model to analyze data structured in classes, to determine their representativity and classification. The model includes an algorithm integrating three parameters: Informational-Weight, Differential-Weight and Tipicity-Contrast. In application we analyze clinical data on 160 patients with lip and palate malformations. The model allows to assess how representative the sample is, using the variables of the cleft, lip and nose along with some expertly determined comparison criteria. Moreover using the Tipicity-Contrast parameter a supervised classification was achieved and has been able to classify correctly, in average, a 93% of the patients. As a result this model can provide helpful auxiliary criteria in medical decision-making.

Keywords: Data analysis, classification, tipicity-contrast, cleft and lip palate, informational-weight, differential-weight.

1 Introduction

We present a new application of the logical-combinatorial approach [1] to a data collected of 160 patients with cleft lip and palate malformations from the Pediatric Hospital of Tacubaya of the Health Institute in Mexico City, in order to ensure the representativity of the sample and the efficiency of such mathematical approach based on the physician's knowledge and experience. The analysis resulted in an algorithm integrating three parameters: Informational Weight (IW), Differential Weight (DW) and Tipicity–Contrast (TC) of each patient's description. The IW and DW parameters measure respectively the degrees of similarity and difference of patients, evaluating the representativity of the object in its own class, but also its contrast with respect to the complement of its class. It is the concept of tipicity-contrast. This is a discriminate analysis through a model that integrates some comparison criteria designed jointly with the physician (expert). This also leads us to integrate both IW and DW parameters in a mathematical function defining the TC practically by averaging these two parameters. We also demonstrate that both parameters, IW and DW, can be used

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 613–620, 2014.

as a supervised classification method. We performed a cross-validation with the sample data and it was demonstrated that the repeatability of the results could be achieved with new data.

2 The Mathematical Model

Let $\boldsymbol{O}=\{O_1, O_2, \ldots, O_m\}$ be a finite set of m objects in the universal set U of all objects in consideration. Each object is described in terms of a finite set of n feature-variables or attributes $X=\{x_1, x_2, \ldots, x_n\}$, where each variable x_i, i=1, . . . , n is defined on its domain $M_i=\{m_{i1}, m_{i2}, \ldots\}\cup\{*\}$, where * denotes absence of *information* [2]. The domain is a set of admissible values for the variable x_i may be quantitative, qualitative, fuzzy, or linguistic in the same set or subset of features. A description of an object O is given by the n-tuple $I(O)=(x_1(O),\ldots.x_n(O))$ with the component function or feature mapping x_i: $M\rightarrow M_i$, i=1…n evaluating the feature x_i of the object. A set M_i of admissible values for the feature x_i does not have a priori algebraic, topological or logic structure. The expert (e.g. the surgeon in a clinical study) is greatly involved in the determination of the variables and their admissible values, including any eventual correlation if necessary to improve the efficiency and accuracy of the clinical decision-making.

Definition 1. Let $\omega \subseteq X$ be a *support set*, where $\omega\neq\varnothing$. A *system of support sets* $\Omega = \{\omega_1, \ldots,\omega_s\}$ is a collection of such subsets. By ωO we denote the ω-part of the object O formed by the variables $x_j\in \omega$.

Remark 1.1: The number of support sets is 2^n-1, where n is the cardinal number of the set X. However, the expert determines the necessary system of sets as well as the objects variables according to the case at hands. We assume the universal set U is structured in r proper subsets K_j also called classes but not necessarily disjoint as in the standard mathematical sense, and not necessarily crisp.

Definition 2. The *partial similarity function* β_ω is defined by:

$$\beta_\omega\left(I(O_i), I(O_j)\right)=1-\left(\sum_{x_t\in\omega}\rho_t C_t\left(x_t(O_i), x_t(O_j)\right)\Big/\sum_{\forall t}\rho_t\right), \tag{1}$$

where ω is a support set, ρ_t is the relevance parameter associated to each variable x_t defined by the expert, and C_t is the *comparison criteria* for each variable $x_i \in X$**.**

Remark 2.1: The partial similarity function is nonnegative, and ranged in the real unit interval [0,1]. The logical-combinatorial approach relies chiefly on the expert's knowledge and experience, mainly for the support sets and the system of support sets.

Definition 3. The *Informational Weight* $IW_j(O)$ of the object O in the class K_j is defined as:

$$IW_j(O)=\left(\sum_{\omega\in\Omega}\rho(\omega)\sum_{O_i\in K_j}\beta_\omega(O,O_i)\right)\Big/\left(|K_j|\sum_{\omega\in\Omega}\rho(\omega)\right), \tag{2}$$

where $|K_j|$ represents the cardinality of the class K_j, and $\rho(\omega)$ is the relevance parameter associated to each support set.

The informational weight IW_j of an object refers to its relevance with respect to a specific class K_j either the object own class or the complement of its class [3]. The relevance is given by the measure of its similarity to all the objects in this class. The greater this measure the more similar the object to the other objects in the class.

Remark 3.1: This parameter IW_j as well is always nonnegative, in the real unit interval [0,1], and depends on the support set and the class K_j.

In the same way, one may measure the object differentiation with respect to the other classes, complement of the class K in the universe of all the objects. This measure, obtained by a discriminant-type analysis, is the object's differential weight *(DW)*. The more dissimilar an object to objects in the remaining classes, the greater the differential weight.

Definition 4. The *Differentiated Weight* $DW_j(O)$ [3] of the object O in the class K_j is defined as:

$$DW_j(O) = \left(\sum_{\omega \in \Omega} \rho(\omega) \sum_{O_i \in K_j} 1 - \beta_\omega(O, O_i) \right) \Big/ \left(m - |K_j| \sum_{\omega \in \Omega} \rho(\omega) \right), \tag{3}$$

where m is the number of objects in the set M.

Remark 4.1: The DW_j is nonnegative and ranges in [0.1]. For a given class, the parameters IW and DW may be considered separately yielding different kinds of classification however relevant, as we will see in the case of the IW.

An object is typical for a class K_j when it is similar to objects in the class K_j with a higher IW value, and the object is dissimilar with objects in the remaining classes with a higher DW value. An object may be representative for more than one class having a great IW for its own class, and a very low DW for the other classes. The higher the TC in a class the more representative the object is for this class.

Definition 5. The *Tipicity – Contrast* $TC_j(O)$ of the object O in the class K_j is defined by the formula:

$$TC_j(O) = \left(IW_j(O) + DW_j(O) \right) / 2 \tag{4}$$

The TC parameter with its IW and DW components is designed to evaluate the membership relationship for an object with respect to the classes that structure the universal set of objects.

3 Tipicity - Contrast Algorithm (TCA)

Given a sample **O** of m objects structured in r classes, described in terms of n feature-variables, with well-defined comparison criteria, the Tipicity–Contrast Algorithm *TCA* consists of the seven steps showed in Fig. 1. Steps 2 to 5 are related with data analysis. Adding step 6, we get a classification method.

4 The Clinical Problem and Its Mathematical Model

The clinical problem consists of congenital malformations in the lip and/or palate (Fig. 2), which are called cleft-primary palate and/or cleft-secondary palate respectively. Primary palate is formed by the prolabium, the premaxilla, and columella. This is the "visible" part of these kinds of malformations. The secondary palate begins at the incisive foramen and extends posteriorly. It includes the horizontal portion of the premaxilla, horizontal portion of the palatine bones, and soft palate [4].

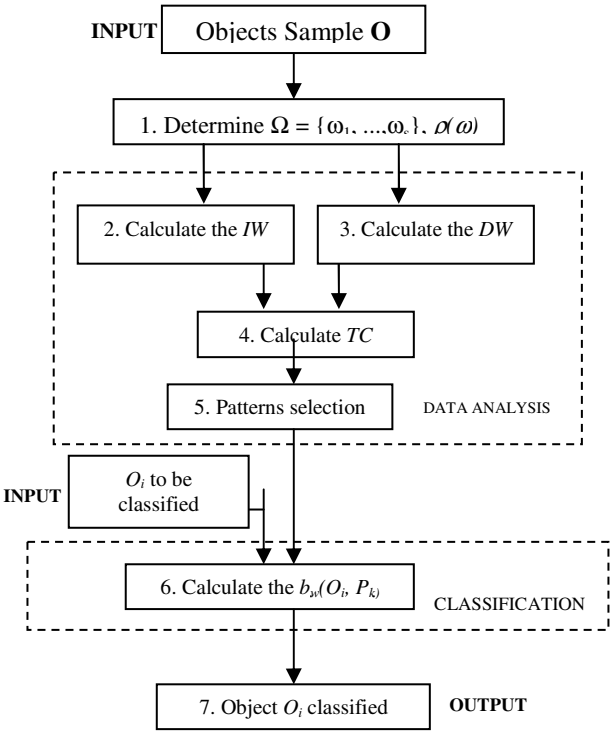

Fig. 1. Block diagram of proposed method

The proper application of the model described above requires the identification of the features-variables and their admissible values as well as the classes and the comparison criteria.

Variables. In order to describe the type of cleft it was necessary to define, in conjunction with the surgeon, the variables related to the different anatomical structures affected (cleft, lip and nose). In this sense, eighteen variables were defined for the description of the patient [5]. Likewise, the comparison criterion for each variable was modeled. All comparison criteria are of *difference*. The minimum value of its domain means that the compared values are equal (there is no difference), and the maximum value means that the compared values are different.

Cleft. Two variables (x_i) $i=1,2$ were defined to describe the cleft: 1) primary palate and 2) secondary palate (Table 1).

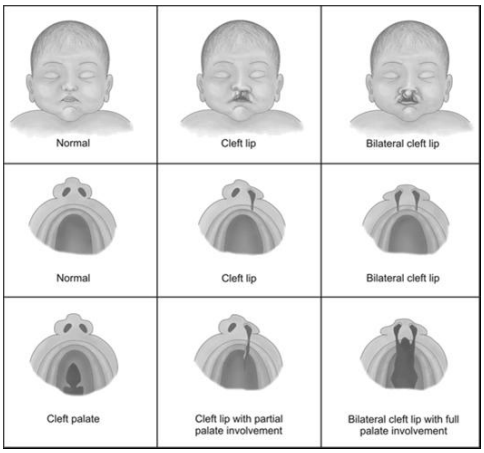

Fig. 2. Cleft lip and palate (unilateral and bilateral) [6]

Table 1. Cleft variables and their comparison criteria [5]

Cleft variables	Comparison criterion
x_1. Primary palate (left and/or right)	$C_i = \frac{\lvert x-y \rvert}{100}$
x_2. Secondary palate (left and/or right)	$C_i = \frac{\lvert x-y \rvert}{55}$

The clefts of the primary palate can be presented in a *unilateral* way (left or right), or in a *bilateral* way (Fig. 2). The latter are formed from the combination of two unilateral fissures. These malformations can have different characteristics with a direct consequence on surgical complexity. For this reason it was necessary to assign a relevance parameter (ρ) to the different clefts. For primary palate, this parameter is in the interval [0, 100] whereas for secondary palate it is in [0, 55].

Lip. In this case, nine variables were defined. These variables have the same 4-valued domain (yes, almost, barely, no). These variable comparison criteria are of the fuzzy type. The physician determined that the difference between two variable values is given by 0.33**d*, where *d*=0,1,2,3 is the distance between them (Table 2).

Nose. In this case, seven variables with different domains were defined, as well as three different fuzzy comparison criteria. As in the lip case each criterion has a homogeneous scale and it is represented by a comparison matrix (Table 3).

Similarity function for cleft palate. This function is defined taking into account the partial similarity (Definition 2), in relation with the different structures considered in patient evaluation. Three support sets were defined, each one corresponding to the cleft, the lip and the nose variables. From Definition 1 we denote

$\Omega = \{\omega_{cleft}, \omega_{lip}, \omega_{nose}\}$ the system of support sets, where $\omega_{cleft} = \{x_1, x_2\}$, $\omega_{lip} = \{x_3, \ldots, x_{11}\}$ and $\omega_{nose} = \{x_{12}, \ldots, x_{18}\}$, and 0.60, 0.20 and 0.20 are the relevance parameters respectively.

Table 2. Lip variables and their comparison criterion

Lip Variables	**Comparison criterion**				
x_3. Symmetry of lip height					
x_4. Normal lip height		yes	almost	barely	No
x_5. Muscular integrity	yes	0	0.33	0.66	1.0
x_6. Skin integrity	almost		0	0.33	0.66
x_7. Mucous membrane integrity	barely			0	0.33
x_8. Symmetry of lip thickness	no				0
x_9. Symmetry of philtral ridges					
x_{10}. Normal sulcus depth					
x_{11}. Presence of cupid arch					

Table 3. Nose variables and their comparison criteria

Nose variables	Comparison criterion				
x_{12}. Symmetry of nasal floor		yes	almost	barely	no
x_{13}. Symmetry of nostril arches	yes	0	0.33	0.66	1
x_{14}. Symmetry of nostrils (vertical plane)	almost		0	0.33	0.66
x_{15}. Symmetry of nostrils (anteroposterior plane)	barely			0	0.33
x_{16}. Nasal septum deviation	no				0

		norm	almost	barely	absent
x_{17}. Length of columella	norm	0	0.33	0.66	1
	almost		0	0.33	0.66
	barely			0	0.33
	absent				0

		greater	normal	Smaller
x_{18}. Width of nasal base	greater	0	0.5	1
	normal		0	0.5
	smaller			0

Definition 6. The partial similarity function β_{cleft} for cleft is given by:

$$\beta_{cleft}(I(P_1), I(P_2)) = 1 - \sum_{t=1}^{2} \rho_t C_t(x_t(P_1), x_t(P_2)), \tag{5}$$

where $\rho_t = \{0.65, 0.35.\}$, t=1,2.

Definition 7. For lip, let β_{lip} be the partial similarity function given by:

$$\beta_{lip}(I(P_1), I(P_2)) = 1 - \sum_{t=3}^{11} \rho_t C_t(x_t(P_1), x_t(P_2)), \tag{6}$$

where $\rho_t = \{0.16, 0.15, 0.14, 0.15, 0.08, 0.12, 0.10, 0.05, 0.05\}$, t=3,…,11.

Definition 8. For nose, let β_{nose} be the partial similarity function defined by:

$$\beta_{nose}(I(P_1), I(P_2)) = 1 - \sum_{t=12}^{18} \rho_t C_t(x_t(P_1), x_t(P_2)), \tag{7}$$

where $\rho_t = \{0.17, 0.25, 0.10, 0.10, 0.11, 0.15, 0.12\}$, t=12,…,18.

5 Sample Description

The *TCA* was applied to a sample of 160 patients with cleft lip and palate grouped into three classes in order to analyze the data collected and determining if it is a representative sample and the validation of the classification method. First class K_1 (excellent), has 40 patients with secondary cleft palate. Here the lip and nose descriptions of these patients have a normal condition (Fig 2.I). The second class K_2 (Very good) is formed by 70 patients, and the class K_3 (Good) is formed by 50 patients. In K_2, and K_3 patients have clefts in both palates (Fig. 2.II and 2.III).

The sample was randomly divided into 80% used as the training set (32 objects in class 1, 56 in class 2 and 40 in class 3) and 20% as the testing set (8, 14, 10 in class 1, 2 and 3 respectively). We made a *5*-fold cross-validation to estimate how good generalization can be made by the TCA. The data was randomly divided to five mutually exclusive subsets and the TCA algorithm was trained and tested five times. In each case, one of the folds was taken as test data and the remaining folds were added to form training data, considering the three classes described above. Thus, five different test results exist one for each training-test configuration.

6 Results

Data analysis. The values of the parameters *IW* and *DW* for each object in the five different test samples were calculated. In all cases the greatest *IW* was obtained precisely in the class to which belongs the object. On the other hand, the parameter *DW* shows the contrast of the object with respect to the classes it does not belong to. Based on these results obtained for the IW and DW, we can conclude that the objects are representative for their own class. Therefore for data analysis, both parameters are useful in determining objects representatively of their class.

Classification. Recall that *TC* is a composite parameter that associates the object *IW* in its class and the difference with the remaining classes by *DW*, the parameter *TC* could be used for classification. This hypothesis was tested by a five folds cross validation, in order to avoid bias in classification. Classification results for the five training sets configuration are summarized in Table 4. All correctly recognised patient

represents the true positive subset (TP), and all patients misclassified represents the false positive subset (FP). The TCA performance was defined as the total number of true positives minus the total number of false positives, divided by the total number of objects, for each training set.

Table 4. TCA performance in the 5 training sets configurations

Trainnig Set	*TP*	*FP*	*Performance*
1	154	6	92.5 %
2	156	4	95%
3	154	6	92.5%
4	153	7	91.25%
5	155	5	93.75%

7 Conclusions

We presented an algorithm for data analysis, based on the concepts of Informational Weight, Differentiated Weight and Tipicity-Contrast of objects taken from a sample structured in a finite number of classes. The sample consists of 160 patients with cleft lip and palate grouped in three classes. With the variables and comparison criteria for the cleft, the lip, and the nose, the IW and DW enabled us to determine the representativity of each patient with respect to each class. Equally important was the collateral result obtained by the Tipicity-Contrast of an object. We classified patients with an efficiency of 93% in average. The performance of the TC was evaluated based on the true positives and false positives. We observed that patients misclassified were at the boundary of two classes, meaning these patients could be in either class according to the expert educated preference. This establishes the TC as a good parameter for supervised classification.

References

1. Martínez-Trinidad, J.F., Guzmán-Arenas, A.: The logical combinatorial approach to pattern recognition, an overview through selected works. Pattern Recogn. 34(4), 741–751 (2001)
2. Lazo-Cortés, M., Ruiz-Shulcloper, J.: Determining the feature relevance for non-classically described objects and a new algorithm to compute typical fuzzy testors. Pattern Recogn. Lett. 16, 1259–1265 (1995)
3. Vega-Alvarado, L., Ortiz Posadas, M.: Análisis de una muestra de pacientes con labio-paladar hendido usando un algoritmo de tipicidad y contraste. Memorias del II Congreso Latinoamericano de Ingeniería Biomédica, La Habana, Cuba (2001)
4. Ortiz-Posadas, M.R., Vega-Alvarado, L., Maya-Behar, J.: A new approach to classify cleft lip and palate. Cleft Palate-Cran. J. 38(6), 545–550 (2001)
5. Ortiz-Posadas, M.R., Vega-Alvarado, L., Toni, B.A.: similarity function to evaluate the orthodontic condition in patients with cleft lip and palate. Med. Hypotheses 63(1), 35–41 (2004)
6. Children's Hospital of Wisconsin, Cleft lip and/or palate. Consulted (February 2009), `http://www.chw.org/display/PPF/DocID/35472/Nav/1/router.asp`