# CLASSICATION BY SUCCESSIVE NEIGHBORHOOD

David Grosser, Henri Ralambondrainy and Noel Conruyt
*Laboratoire d'Informatique et Mathématiques, Université de la Réunion, 97490 Sainte-Clotilde, France*

Keywords: Classification, Similarity, Nearest neighbors, Structured data, Systematics.

Abstract: Formalization of scientific knowledge in life sciences by experts in biology or Systematics produces arborescent representations whose values could be present, absent or unknown. To improve the robustness of the classification process of those complex objects, often partially described, we propose a new classification method which is iterative, interactive and semi-directed. It combines inductive techniques for the choice of discriminating variables and search for nearest neighbors based on various similarity measures which take into account structures and values of the objects for the neighborhood computation.

## 1 INTRODUCTION

Systematics is the scientific discipline that deals with listing, describing, naming, classifying and identifying living beings. In the frame of environmental sciences, the acquisition and production of knowledge on biological specimens and taxa is an essential part of the work of systematicians (Winston, 1999). Indeed, being able to describe, classify and identify a specimen from morphological characters is a first step for monitoring biodiversity because it gives access to information relative to its species name (Biology, Geography, Ecology, etc.). This process can be assisted with computer science decision support tools. In return, such complex domains deliver interesting symbolical and numerical knowledge representation and processing problems to the knowledge engineering and computer science community.

Indeed, classical discrimination methods developed in the frame of data analysis or machine learning, such as classification or decision trees (Breiman et al., 1984), (Quinlan, 1986) or more recent methods developed in the data mining field such as association rules mining (Piatetsky-Shapiro, 1991) or Multifactor dimensionality reduction (Zhu and Davidson, 2007) are not sufficient, because they do not cope with relations between attributes, missing data, and are not very tolerant to errors in descriptions.

The considered problem that we are faced with is to determine the class of a structured description that is partially answered and, eventually contains errors, from a referenced case base, this last one be a priori classified by qualified experts in k-classes. The proposed discrimination method proceeds by inference of successive neighboring. It is inductive, interactive, iterative and semi-directed. It combines inductive techniques of discriminatory variables and neighbors search, with the help of a similarity measure that takes into account the structure (dependencies of variables) and the content (missing and unknown values).

## 2 DATA REPRESENTATION

Within a knowledge base, observations are described with the help of *descriptive models*. A descriptive model represent an ontological knowledge about the considered domain and contains descriptors structure and organization.

### 2.1 The Descriptive Model

The descriptive model (fig. 1), or schema, is a rooted tree $\mathcal{M} = (\mathcal{A}, \mathcal{U})$, where $\mathcal{A}$ is a set of nodes(attributes), and $\mathcal{U}$ a set of edges. Leaves are single classical attributes, as numerical or nominal ones, called "basic attributes". Nodes are "structured attributes", sub-trees made of several attributes. For example, $A_j :< A_1,\ldots,A_p >$ denotes a structured attribute where $A_j$ is the root of the sub-tree and $A_1,\ldots,A_p$, are the sons, structured or basic attributes, the components of $A_j$.

Grosser D., Ralambondrainy H. and Conruyt N. (2009).
CLASSIFICATION BY SUCCESSIVE NEIGHBORHOOD.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 288-291
DOI: 10.5220/0002299002880291

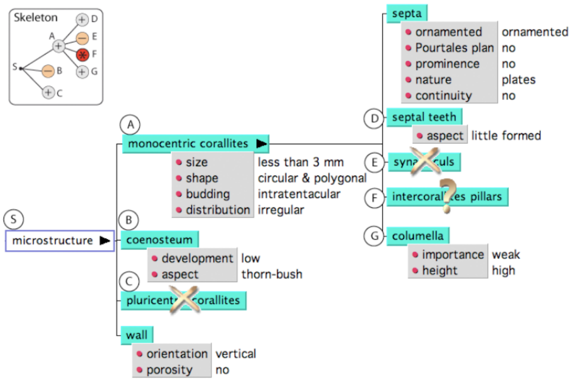

Figure 1: This example shows a description part of a specimen belonging to the genus *Stylocoeniella*.

## 2.2 Object Representation

An object, described by a schema $\mathcal{M}$, is a tree derived from $\mathcal{M}$ where each basic attribute has been valued (fig. 1). A *skeleton* represents the structure of an object, it describes the state of each component: present (+), absent (or missing) (-), or unknown ($*$). We denote by $S = \{+,-,*\}$. A map $\sigma : \mathcal{A} \rightarrow S$ defines a labeled rooted tree $H_\sigma = (\mathcal{A}_\sigma, \mathcal{U})$ where $\mathcal{A}_\sigma = \{(A_j, \sigma(A_j)) | A_j \in \mathcal{A}\}$. The skeleton of an object is represented by $H_\sigma$. The nodes of a skeleton must respect the following consistency constraints: for each structured attribute $B :< B_l >_{l \in L}$, we must have:

1. "The sons of a missing node must be missing": if $\sigma(B) = -$ then $\sigma(B_l) = -$, for $l \in L$,
2. "The sons of an unknown node are unknown or missing": if $\sigma(B) = *$ then $\sigma(B_l) = *|-$, for $l \in L$.
3. "The sons of a present node may be present, missing or unknown": if $\sigma(B) = +$ then $\sigma(B_l) = +|-|*$, for $l \in L$.

We denote by $\mathcal{H}$ the set of skeletons that are consistent. Assume that is given a set of basic attributes names $A_q$ and corresponding domains $D_q$ for $q \in Q$. For any object, a basic attribute $A_q$ is valued in $D_q$ only If the attribute is present. Missing values will be denoted by $\perp$ and unknown values by $*$, then $\Gamma_q = D_q \cup \{\perp\} \cup \{*\}\}$ is the new domain for a basic attribute $A_q$. The set of values of an object $o$ is $v_o = (v_q)_{q \in Q}$ where $v_q \in \Gamma_q$. Let $\Gamma_Q = \sqcap_{q \in Q} \Gamma_q$, an object is described with its skeleton and values : $o = (H_{\sigma_o}, v_o) \in \mathcal{E} = \mathcal{H} \times \Gamma$.

# 3 SIMILARITY MEASURES

In this section, we propose a function to evaluate the similarity level between a pair of skeletons. The weight $m(A_j)$ of the attribute $A_j$ is the number of attributes which is made of. It means that the contribution of a structured attribute in the similarity evaluation of skeletons depends on the number of its components. Given a comparison map $\lambda : S \times S \longrightarrow [0,1]$ (table 1), the structural weighted similarity measure, between the skeletons $H_{\sigma_1}$ et $H_{\sigma_2}$, is defined as:

$$\zeta_{SW}(H_{\sigma_1}, H_{\sigma_2}) = \frac{\Sigma_{j \in J} m(A_j) \lambda(\{\sigma_1(A_j), \sigma_2(A_j)\})}{\Sigma_{j \in J} m(A_j)}.$$

If we take $\alpha_j^1 = \beta_j^1 = \alpha_j^2 = \gamma_j = \beta_j^2 = 0$ (table 1) and a weight equal to 1 for all attributes, then the proposed function is the well-known Sokal index (Sneath and Sokal, 1973). Another possible choice is to take $\beta_j^1 = \beta_j^2 = \gamma_j = 0$ for the comparison of unknown values and define $\lambda(\{\sigma_1(A_j) = +, \sigma_2(A_j) = -\}) = \alpha_j^1$ as the ratio of missing sons of $A_j$ in the skeleton $H_{\sigma_1}$.

Table 1: Values of a comparaison map $\lambda$ for a given node $A_j$.

| $H_{\sigma_1} \diagdown H_{\sigma_2}$ | $+$ | $-$ | $*$ |
|---|---|---|---|
| $+$ | 1 | $\alpha_j^1$ | $\beta_j^1$ |
| $-$ | $\alpha_j^2$ | 1 | $\gamma_j$ |
| $*$ | $\beta_j^2$ | $\gamma_j$ | 1 |

For some applications, users wish emphasize present rather than missing or unknown attributes. For that purpose, we define the structural recursive similarity measure built from a map $\lambda_r$ as follow: if $A_j$ is a basic attribute, presents both in $H_{\sigma_1}$ and $H_{\sigma_2} : \sigma_1(A_j) = \sigma_2(A_j) = +$ then:

$$\lambda_r(+,+) = 1, \qquad (1)$$

else $A_j :< A_l >_{l \in L}$ is a structured attribute and

$$\lambda_r(+,+) = \frac{1 + \Sigma_l m(A_l) \lambda_r(\{\sigma_1(A_l), \sigma_2(A_l)\})}{m(A_j)} \qquad (2)$$

This similarity measure is recursively computed from the root $A$:

$$\zeta_{SR}(H_{\sigma_1}, H_{\sigma_2}) = \lambda_r(\{\sigma_1(A), \sigma_2(A)\}) \qquad (3)$$

# 4 CLASSIFICATION BY SUCCESSIVE NEIGHBORHOOD

The proposed classification method allows to determine the membership of an individual to a particular class which is partially described by a user and comprising eventually some errors. It takes into account dependences relations between attributes. The principle consists with to select a neighborhood $V$, i.e. a descriptions set (individuals or classes) close to current description with the help of the similarity measure. A set of candidates classes is computed from the neighbor set. The method seeks then to supplement the missing information, firstly with the application of coherency rules, secondly by proposing a set of discriminants attributes. A new neighborhood is then computed on the basis of the new partial description. The process is reiterates until obtaining a homogeneous descriptions set.

## 4.1 Neighborhood Classification Algorithm

The iterative process to predict the class of a specimen, from a given description $e$, is made of the following steps:

1. Initialize radius value $\Delta$ to the max of the distance of $e$ to the set of observations.
2. Determine the set of objects inside of the sphere of radius $\Delta$ centered at $e$,
3. Compute the classification scores of the *a priori* classes,
4. Compute a new radius value $\Delta$ from the set of neighbors,
5. Repeat 2, 3, 4 until stopping condition is satisfied

### 4.1.1 Neighbors Set

The neighbors of $e$ at iteration $m$ is the set of objects inside of the sphere of radius $\Delta_m$ centered at $e$:

$$N_{(m)} = \{o \in O \mid d(e,o) < \Delta_m\}.$$

The radius value is determined from the maximum distance, a dissimilarity measure between $e$ and a set $A$ :

$$D_{max}(e,A) = max_{a \in A} d(e,a)$$

then $\Delta_m$ is written: $\Delta_m = D_{max}(e, N_{(m-1)})$. It is easy to show that $\{\Delta_m\}$ is a decreasing sequence. If each object $o_i$ has a normalized weight $p_i$, then the dispersion distance:

$$D_{disp}(e,A) = \Sigma_{o_i \in A} p_i d^2(e,o_i)$$

is suitable, we can show that $\{\Delta_m = D_{disp}(e, N_{(m-1)})\}$ is still a decreasing sequence.

### 4.1.2 Class Classification Score

The examples $O$ are pre-classified into classes denoted by $\{C_l\}_{l \in K}$. Let $Pr(C_l|N_{(m)}) = \frac{|C_l \cap N_{(m)}|}{|N_{(m)}|}$ be the probability of the cluster $C_l$ given the set of neighbors $N_{(m)}$ or the relative frequency of the cluster $C_l$ in the set of neighbors $N_{(m)}$. The label that should be assigned to input $e$ will be chosen from clusters such as its probability $Pr(C_l|N_m)$ are significantly different from the prior probability of the cluster $Pr(C_l|O) = \frac{|C_l|}{|0|}$. Usual significant statistical test of frequencies may be used or defined user threshold for this purpose. Then, the classification score of the class $C_l$ at iteration $m$ is

$$R_l = \frac{Pr(C_l|N_m)}{Pr(C_l|O)}.$$

### 4.1.3 Stopping Condition

The initial value of the radius is $\Delta_0 = D_{max}(e,O)$. As the sequence of radius is decreasing, then the sequence of the set of neighbors $\{N_{(m)}\}$ is also decreasing. The iterative process is suspended, when the maximum of a classification score $R_l$ of a class $C_l$ is greater than a threshold $r_0 > 1$. In practice $r_0 = 2$, the class with the best score is proposed to the user as the label of the specimen. Otherwise, a minimum size of the neighbors set (ten percent of the population size for example) is used as a stopping criteria. We can notice that it is more easier to the user to fix the previous thresholds than to give a good number $k$ of nearest neighbors.

### 4.1.4 Discriminant Attributes Selection

An ordered list of informative variables is computed at each iteration of the classification process. The first element is exposed as a question to the user who can choose an alternative variable from the list or an unknown answer. The list is built in function of several criteria:

1. Attribute potentiality. The method considers only at each step the attributes which can be indicated, i.e., those for which there exists a components (nodes) chain whose presence is proven.
2. Discriminant power. Choice of different classical criteria computing the information gain used in machine learning such as Shannon entropy or Gini index.
3. Background knowledge by using attributes weighting in the descriptive model.

# 5 ILLUSTRATIVE EXAMPLE

The famous Anderson's Iris data is used to illustrate the neighborhood classification algorithm. This data set contains 150 plants from 3 specific species characterized by 4 attributes. The task is to predict the class of a given specimen, its description $e$ is a slight modification of a plant description of a *setoza* specy (see Table 2).

Table 2: Specimen.

| Attribute | petal | | sepal | |
|---|---|---|---|---|
| Specimen | width | length | width | length |
| $e_1$ | - | 3.1 | - | - |
| $e_2$ | - | 3.1 | - | 2.19 |
| $e$ | 7 | 3.1 | 5.97 | 2.19 |

To test the performance of the algorithm, we start to identify the given specimen, using only the *petal length value*, the other values are considered as "unknown" (specimen $e_1$ in Table 2). After 16 iterations (Table 3), the algorithm stops as the size of neighbors is less than 15 (the threshold has been fixed to be ten percent of the population size). Scores of the 3 classes are not satisfactory as they are lower than 2. The attribute *sepal length* is proposed to be assigned a value, according to its discriminatory ability. The algorithm is applied to the resulted specimen $e_2$. After 46 iterations, the class setoza has the best score greater than 2, and is proposed to be the specimen label. In contrast to decision tree methods, we can notice that the identification process can be performed with attributes that have not necessarily a good discriminatory ability.

Table 3: Experiments.

| **Cluster** | **versicolor** | **virginica** | **setoza** |
|---|---|---|---|
| Size *Cluster* (C) | 50 | 50 | 50 |
| $Pr(C\|O)$ | 33.3 | 33.3 | 33.3 |
| *Identification of* $e_1$: **versicolor** ? *(iteration =16)* | | | |
| Size *Neighbor* (N) | 5 | 3 | 4 |
| $Pr(C\|N)$ | 41.7 | 25 | 33.3 |
| *Cluster score R* | 1.25 | 0.75 | 1 |
| *Identification of* $e_2$: **setoza** *(iteration =46)* | | | |
| Size N*eighbor* | 0 | 21 | 48 |
| $Pr(C\|N)$ | 0 | 30.4 | 69.6 |
| *Cluster score R* | 0 | 0.91 | 2.08 |
| *Identification of* $e$: **setoza** *(iteration =66)* | | | |
| Size N*eighbor* | 0 | 24 | 49 |
| $Pr(C\|N)$ | 0 | 32.9 | 67.1 |
| *Cluster score R* | 0 | 0.99 | 2.01 |

# 6 CONCLUSIONS AND PERSPECTIVES

To identify a biological object and to associate a taxon to it, most of the time systematicians proceed in two phases. The synthetic phase, by global observation of the most visible characters reduces the field of investigation. The analytical phase, by precise observation of discriminating attributes refines research until obtaining the result. The classification by successive neighborhood from a partial description that we propose presents the interest to correspond to the reasoning followed by biologists. Starting from a partial description generally containing the most visible or easy to observe and describe attributes, the method suggests relevant information necessary to supplement to determine the most probable class. Moreover, it is error tolerant because an erroneous information can nevertheless lead to a satisfactory result due to the fact that a smooth matching is carried using similarity function out on filled values rather than a strict one.

We expect that the method is generic and applicable on any fields where structured or semi-structured data are considered, such as XML data format, RDF graph structures or OWL Ontologies. It's enough to lay out an operator of generalization and a similarity index adapted to the considered data. This method is in evaluation progress on the " knowledge base on corals of the Mascareignes archipelago" which counts approximately 150 taxa and 800 complex descriptions.

# REFERENCES

Breiman, L., Freidman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth, Belmont.

Piatetsky-Shapiro, G. (1991). *Discovery, analysis, and presentation of strong rules*. Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Sneath, H. and Sokal, R. (1973). *Numerical Taxonomy*. W.H. Freeman.

Winston, J. E. (1999). *Describing Species: Practical Taxonomic Procedure for Biologists*. New York: Columbia University Press.

Zhu, X. and Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*. Idea Group Inc.