# Stereotype Extraction with Default Clustering

**Julien Velcin and Jean-Gabriel Ganascia**
LIP6, Université Paris VI
8 rue du Capitaine Scott
75015 PARIS, FRANCE

## Abstract

The concept of stereotype seems to be really adapted when wishing to extract meaningful descriptions from data, especially when there is a high rate of missing values. This paper proposes a logical framework called *default clustering* based on default reasoning and local search techniques. The first experiment deals with the rediscovering of initial descriptions from artificial data sets, the second one extracts stereotypes of politicians in a real case generated from newspaper articles. It is shown that default clustering is more adapted in this context than the three classical clusterers considered.

## Introduction

Conceptual clustering [Michalski, 1980] is a fundamental machine learning task that is applied in various areas such as image analysis, analytical chemistry, biology, sociology. It takes a set of object descriptions as input and creates a classification scheme. The conceptual descriptions of clusters are of particular interest to reason about the categories themselves, to compare different data sets and to predict new observations. This work focuses on the extraction of such conceptual descriptions, especially in the specific context of missing data.

Automatic inductive techniques have to deal with missing information, due to voluntary omissions, human error, broken equipment [Newgard and Lewis, 2002]. In the context of sparse data, i.e. with a huge amount of missing values, the concept of stereotype seems more appropriate than the usual one of prototype to describe data clusters. Therefore, our goal is to extract stereotype sets that represent the data sets as well as possible. By analogy to default logic [Reiter, 1980], which is a specific logic for default deduction, we make use of *default subsumption*, which is a specific logic for default induction, to build such stereotypes.

Section 1 presents a new approach to conceptual clustering when missing information exists. Section 2 proposes a general framework in the attribute-value formalism. The new notion of default subsumption is introduced, before seeing how the concept of stereotype makes it possible to name clusters. A stereotype set extraction algorithm based on local search techniques is then presented. Section 3 concerns experiments, first on artificial data sets and secondly with a real data case generated from newspaper articles.

## 1 Conceptual clustering with sparse data

### 1.1 Dealing with missing values

This paper proposes a clustering method that deals with high rates of missing values. But contrary to algorithms such as k-modes (categorical version of k-means) or EM, that can easily lead to local optima, we have chosen to achieve the clustering using a combinatorial optimization approach, like in [Figueroa *et al.*, 2003] or [Sarkar and Leong, 2001]. Note that our goal is not only to cluster examples but also and mainly to describe the cluster in a way that is simple and easy to understand. The problem can thus be stated as finding readable, understandable, consistent and rich descriptions from the data.

### 1.2 Overview of default logic

During the eighties, there were many attempts to model deductive reasoning when missing information exists. A lot of formalisms were developed to encompass the inherent difficulties of such models, especially their non-monotony: close-world assumption, circumscription, default logic, etc. Since our goal is to deal with missing values, it seems natural to take advantage of this work. The default logic formalism, introduced by R. Reiter in 1980, was chosen because it seemed to correspond well to our problem.

This logic for default reasoning is based on the notion of default rule, through which it is possible to infer new formulas when the hypotheses are not inconsistent with the current context. More generally, a default rule always has the following form: $A : B_1, B_2 \ldots B_n / C$ where $A$ is called the prerequisite, $B_i$ the justifications and $C$ the conclusion. This default rule can be interpreted as follows: if $A$ is known to be true and if it is consistent to assume $B_1, B_2 \ldots B_n$ then conclude $C$. For instance, let us consider the default rule below related to the experiments of the last section:

$$\frac{politician(X) \wedge introducedAbroad(X) : \neg diplomat(X)}{traitor(X)}$$

This rule translates an usual way of reasoning for many people living in France at the end of the 19th century. It states that the conclusion `traitor(X)` can be derived if `X` is a politician who is known to be introduced abroad and that we cannot prove that he is a diplomat.

The key idea here is to use similar observations and their descriptions to infer new information instead of default rules, but the underlying mechanism is the same. The following subsection explains the transition from default logic to default induction.

### 1.3 Default clustering

E. Rosch saw the categorization itself as one of the most important issues in cognitive science [Rosch, 1975]. She introduced the concept of prototype as the ideal member of a category. Whereas categorization makes similar observations fit together and dissimilar observations be well separated, clustering is the induction process in data mining that actually build such categories. More specifically, conceptual clustering is a machine learning task defined by R. Michalski [Michalski, 1980] which does not require a teacher and uses an evaluation function to discover classes that have appropriate conceptual descriptions. Conceptual clustering was principally studied in a probabilistic context (see, for instance, D. Fisher's Cobweb algorithm [Fisher, 1987]) and rarely on really sparse data sets. For instance, the experiments done by P.H. Gennari do not exceed 30% of missing values [Gennari, 1990].

As seen above, default logic is a logic for deduction depending on background knowledge. This paper proposes a new technique called *default clustering* which uses a similar principle but for induction when missing information exists. The main assumption is the following: if an observation is grouped with other similar observations, you can use these observations to complete unknown information in the original fact if it remains consistent with the current context. Whereas default logic needs implicit knowledge expressed by default rules, default clustering only uses information available in the data set. The next section presents this new framework. It shows how to extract stereotype sets from very sparse data: first it extends classical subsumption (see section 2.1), next it discusses stereotype choice (see section 2.2), and finally it proposes a local search strategy to find the best solution (see section 2.4).

## 2 Logical framework

This section presents the logical framework of default clustering in the attribute-value formalism (an adaptation to conceptual graphs can be found in [Ganascia and Velcin, 2004]). The description space is noted $\mathcal{D}$, the descriptor space (i.e. the values the attributes can take) $\mathcal{V}$ and the example set $E$. The function $\delta$ maps each example $e \in E$ to its description $\delta(e) \in \mathcal{D}$.

### 2.1 Default subsumption

Contrary to default logic, the problem here is not to deduce, but to induce knowledge from data sets in which most of the information is unknown. Therefore, we put forward the notion of *default subsumption*, which is the equivalent for subsumption of the default rule for deduction. Saying that a description $d \in \mathcal{D}$ subsumes $d' \in \mathcal{D}$ by default means that there exists an implicit description $d''$ such that $d'$ completed with $d''$, i.e. $d' \wedge d''$, is more specific than $d$ in the classical sense, which signifies that $d' \wedge d''$ entails $d$. The exact definition follows:

**Definition 1** *$d$ subsumes $d'$ by default (noted $d \leq_D d'$) iff $\exists\ d_c$ such that $d_c \neq \perp$ and $d \leq d_c$ and $d' \leq d_c$ where $t \leq t'$ stands for $t$ subsumes $t'$ in the classical sense. $d_c$ is a minorant of $d$ and $d'$ in the subsumption lattice.*

To illustrate our definition, here are some descriptions based on binary attributes that can be compared with respect to the default subsumption:

$d_1 = \{(Traitor = yes), (Internationalist = yes)\}$
$d_2 = \{(Traitor = yes), (Connection\_with\_jews = yes)\}$
$d_3 = \{(Patriot = yes)\}$

$d_1 \leq_D d_2$ and $d_2 \leq_D d_1$ because $\exists\ d_c$ such that $d_1 \leq d_c$ and $d_2 \leq d_c$: $d_c = \{(Traitor = yes), (Internationalist = yes), (Connection\_with\_jews = yes)\}$.

However, considering that a patriot cannot be an internationalist and vice-versa, i.e. ¬((Patriot=yes) ∧ (Internationalist=yes)), which was an implicit statement for many people living in France at the end of the 19th century, $d_1$ does not subsume $d_3$ by default, i.e. $\neg(d_1 \leq_D d_3)$.

*Property 1* The notion of default subsumption is more general than classical subsumption since, if $d$ subsumes $d'$, i.e. $d \leq d'$, then $d$ subsumes $d'$ by default, i.e. $d \leq_D d'$. The converse is not true.

*Property 2* The default subsumption relationship is symmetrical, i.e. $\forall d\ \forall d'$ if $d \leq_D d'$ then $d' \leq_D d$.

Note that the notion of default subsumption may appear strange for people accustomed to classical subsumption because of the symmetrical relationship. As a consequence, it does not define an ordering relationship on the description space $\mathcal{D}$. The notation $\leq_D$ may be confusing with respect to this symmetry, but it is relative to the underlying idea of generality.

### 2.2 Concept of stereotype

In the literature of categorization, Rosch introduced the concept of prototype [Rosch, 1975; 1978] inspired by the family resemblance notion of Wittgenstein [Wittgenstein, 1953] (see [Shawver, 2004] for an electronic version and [Narboux, 2001] for an analysis focused on family resemblance). Even if our approach and the original idea behind the concept of prototype have several features in common, we prefer to refer to the older concept of stereotype that was introduced by the publicist W. Lippman [Lippman, 1922]. For him, stereotypes are perceptive schemas (a structured association of characteristic features) shared by a group about other person or object categories. These simplifying and generalizing images about reality affect human behavior and are very subjective. Below are three main reasons to make such a choice.

First of all, the concept of prototype is often misused in data mining techniques. It is reduced to either an average observation of the examples or an artificial description built on the most frequent shared features. Nevertheless, both of them are far from the underlying idea in family resemblance. Especially in the context of sparse data, it seems more correct to speak about a combination of features found in different example descriptions than about average or mode selection. The second argument is that the notion of stereotype is often defined as an imaginary picture that distorts the reality. Our goal is precisely to generate such pictures even if they are caricatural of the observations. Finally, these specific descriptions are better adapted for fast classification (we can even say discrimination) and prediction than prototypes. This last feature is closely linked to Lippman's definition.

In order to avoid ambiguities, we restrict the notion of stereotype to a specific description $d \in \mathcal{D}$ associated to (we can say "covering") a set of descriptions $D \subset \mathcal{D}$. However, the following subsection does not deal just with stereotypes but with stereotype sets to cover a whole description set. The objective is therefore to automatically construct stereotype sets, whereas most of the studies focus on already fixed stereotype usage [Rich, 1979; Amossy and Herschberg Pierrot, 1997]. Keeping this in mind, the space of all the possible stereotype sets is browsed in order to discover the best one, i.e. the set that best covers the examples of $E$ with respect to some similarity measure. But just before addressing the search itself, we should consider both the relation of relative cover and the similarity measure used to build the categorization from stereotype sets.

### 2.3 Stereotype sets and relative cover

Given an example $e$ characterized by its description $d = \delta(e) \in \mathcal{D}$, consider the following statement: the stereotype $s \in \mathcal{D}$ can cover $e$ if and only if $s$ subsumes $d$ by default. It means that in the context of missing data each piece of information is so crucial that even a single contradiction prevents the stereotype from being a correct generalization. Furthermore, since there is no contradiction between this example and its related stereotype, the stereotype may be used to complete the example description.

In order to perform the clustering, a very general similarity measure $M_{sim}$ has been defined, which counts the number of common descriptors of $\mathcal{V}$ belonging to two descriptions, ignores the unknown values and takes into account the default subsumption relationship:

$$\begin{array}{rcl} M_{sim}\colon \mathcal{D}\times\mathcal{D} & \rightarrow & \mathbb{N}^+ \\ (d_i, d_j) & \mapsto & M_{sim}(d_i, d_j) = |\{v \in d / d = d_i \wedge d_j\}| \\ & & \text{if } d_i \leq_D d_j \text{ and} \\ & & M_{sim}(d_i, d_j) = 0 \text{ if } \neg(d_i \leq_D d_j), \end{array}$$

where $d_i \wedge d_j$ is the least minorant of $d_i$ and $d_j$ in the subsumption lattice.

Let us now consider a set $S = \{s_\emptyset, s_1, s_2 \ldots s_n\} \subset \mathcal{D}$ of stereotypes. $s_\emptyset$ is the absurd-stereotype linked to the set $E_\emptyset$. Then, a categorization of $E$ can be calculated using $S$ with an affectation function that we called *relative cover*:

**Definition 2** *The* relative cover *of an example $e \in E$, with respect to a set of stereotypes $S = \{s_\emptyset, s_1, s_2 \ldots s_n\}$, noted $C_S(e)$, is the stereotype $s_i$ if and only if:*

- $s_i \in S,$
- $M_S(\delta(e), s_i) > 0,$
- $\forall k \in [1, n], k \neq i, M_{sim}(\delta(e), s_i) > M_{sim}(\delta(e), s_k).$

It means that an example $e \in E$ is associated to the most similar and "covering-able" stereotype *relative* to the set $S$. If there are two competitive stereotypes with an equal higher score or if there is no covering stereotype, then the example is associated to the absurd-stereotype $s_\emptyset$. In this case, no completion can be calculated for $e$.

### 2.4 Stereotype extraction

In this paper, default reasoning is formalized using the notions of both default subsumption and stereotype set. Up to now, these stereotype sets were supposed to be given. This section shows how the classification can be organized into such sets in a non-supervised learning task. It can be summarized as follows. Given:

1. An example set $E$.
2. A description space $\mathcal{D}$.
3. A description function $\delta$: $E \longrightarrow \mathcal{D}$ which associates a description $\delta(e) \in \mathcal{D}$ to each example belonging to the training set $E$.

The function of a non-supervised learning algorithm is to organize the initial set of individuals $E$ into a structure (for instance a hierarchy, a lattice or a pyramid). In the present case, the structure is limited to partitions of the training set, which corresponds to searching for stereotype sets as discussed above. These partitions may be generated by $(n + 1)$ stereotypes $S = \{s_\emptyset, s_1, s_2 \ldots s_n\}$: it is sufficient to associate to each $s_i$ the set $E_i$ of examples $e$ belonging to $E$ and covered by $s_i$ relative to $S$. The examples that cannot be covered by any stereotype are put into the $E_\emptyset$ cluster and associated to $s_\emptyset$.

To choose from among the numerous possible partitions, which is a combinatorial problem, a non-supervised algorithm requires a function for evaluating stereotype set relevance. Because of the categorical nature of data and the previous definition of relative cover, it appears natural to make use of the similarity measure $M_{sim}$. This is exactly what we do by introducing the following evaluation function $h_E$:

**Definition 3** *$E$ being an example set, $S = \{s_\emptyset, s_1, s_2 \ldots s_n\}$ a stereotype set and $C_S$ the function that associates to each example $e$ its relative cover, i.e. its closest stereotype with respect to $M_{sim}$ and $S$, the evaluation function $h_E$ is defined as follows:*

$$h_E(S) = \sum_{e \in E} M_{sim}(\delta(e), C_S(e))$$

While k-modes and EM algorithms are straightforward, i.e. each step leads to the next one until convergence, we reduce here the non-supervised learning task to an optimization problem. This approach offers several interesting features: avoiding local optima (especially with categorical and sparse data), providing "good" solutions even if not the best ones, better control of the search. In addition, it is not necessary to specify the number of expected stereotypes that is also discovered during the search process.

There are several methods for exploring such a search space (hill-climbing, simulated annealing, etc.). We have chosen the meta-heuristic called *tabu search* which improves the local search algorithm. Remember that the local search process can be schematized as follows: 1. An initial solution $S_{ini}$ is given (for instance at random). 2. A neighborhood $\mathcal{P}$ is calculated from the current solution $S_i$ with the assistance of permitted movements. These movements can be of low influence (enrich one stereotype with a descriptor, remove a descriptor from another) or of high influence (add or retract one stereotype to or from the current stereotype set). 3. The

```
$S_c \leftarrow \{s_\emptyset\}$ ; the current solution
$S_b \leftarrow S_c$ ; the best up-to-now solution
$T \leftarrow \emptyset$ ; the list that contains tabu-attributes
for $i \leftarrow 1$ to $NStep$ do {
    let $S_c$ be $\{s_\emptyset, s_1, s_2 \ldots s_k\}$
    $\mathcal{P} \leftarrow \emptyset$ ; initialize the neighborhood
    for all $A_i \notin T$ do {
        for $m \leftarrow 1$ to $k$ do {
            for all $v \leftarrow (A_i = V_{ij})$ do
            if $v$ does not belong to a stereotype of $S_c$ {
                $S \leftarrow \{s_1 \ldots s_{m-1}, s_m \cup v, s_{m+1} \ldots s_k\}$
                $\mathcal{P} \leftarrow \mathcal{P} \cup S$ }
            if $v$ belongs to a stereotype of $S_c$ {
                $s' \leftarrow s_m \setminus v$
                if $s' \neq \emptyset$ then $S \leftarrow \{s_1 \ldots s_{m-1}, s', s_{m+1} \ldots s_k\}$
                else $S \leftarrow \{s_1 \ldots s_{m-1}, s_{m+1} \ldots s_k\}$
                if $S \neq \emptyset$ then $\mathcal{P} \leftarrow \mathcal{P} \cup S$ }
        }
        for all $v \leftarrow (A_i = V_{ij})$
        if $v$ does not belong to a stereotype of $S_c$ do {
            $s_n \leftarrow \{v\}$
            $S \leftarrow \{s_1, s_2 \ldots s_m, s_n\}$
            $\mathcal{P} \leftarrow \mathcal{P} \cup S$ }
    }
    $S_N \leftarrow$ argmax $_{S \in \mathcal{P}}(h_E(S))$
    $T$ is updated, depending on the chosen attribute $A_i$
        that permits to pass from $S_c$ to $S_N$.
    $S_c \leftarrow S_N$
    if $h_E(S_c) > h_E(S_b)$ then $S_b \leftarrow S_c$ }
return $S_b$
```

Figure 1: The default clustering algorithm

best movement, relative to the evaluation function $h_E$, is chosen and the new current solution $S_{i+1}$ is computed. 4. The process is iterated a specific number of times $NStep$ and the best up-to-now discovered solution is recorded. Then, the solution is the stereotype set $S_b$ that best maximizes $h_E$ in comparison to all the crossed sets.

As in almost all local search techniques, there is a trade-off between exploitation, i.e. choosing the best movement, and exploration, i.e. choosing a non optimal state to reach completely different areas. The tabu search extends the basic local search by manipulating short and long-term memories which are used to avoid loops and to intelligently explore the search space. This meta-heuristic is detailed in [Glover and Laguna, 1997] and its application to clustering can be found in [Al-Sultan, 1995]. Note that only the short-term memory was used at this stage of our work.

## 2.5 Default clustering algorithm

Fig. 1 is the main frame of the default clustering algorithm. It is based on a very basic version of tabu search that tries to maximize our function $h_E$. $A_i$ denotes the $i$th attribute and $E$ is the example set. $S_c$ and $S_b$ stands respectively for the current and for the best solution. $NStep$ is the maximal number of iterations, $\mathcal{P}$ the current neighborhood and $T$ the tabu-list that contains tabu-attributes. If an attribute $A_i$ is in the tabu-list then no descriptor $(A_i = V_{ij})$ can be used to calculate the neighborhood of the current solution.

A "no-redundancy" constraint has been added in order to obtain a perfect separation between the stereotypes. In the context of sparseness, it seems really important to extract contrasted descriptions which are used to quickly classify the examples, as does the concept of stereotype introduced by Lippman.

A new constraint called *cognitive cohesion* is now defined. It verifies cohesion within a cluster, i.e. an example set $E_j \subset E$, relative to the corresponding stereotype $s_j \in S$. Cognitive cohesion is verified if and only if, given two descriptors $v_1$ and $v_2 \in \mathcal{V}$ of $s_j$, it is always possible to find a series of examples that makes it possible to pass by correlation from $v_1$ to $v_2$. Below are two example sets with their covering stereotype. The example on the left verifies the constraint, the one on the right does not.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1 :$ | $a_0$ , | $b_1$ , | $d_5$ , | $f_0$ , | $h_0$ | $s_2 :$ | $a_0$ , | $b_1$ , | $d_5$ , | $f_0$ , | $h_0$ |
| $e_1 :$ | $a_0$, | ?, | ?, | ?, | $h_0$ | $e_0 :$ | $a_0$, | $b_1$, | ?, | ?, | ? |
| $e_2 :$ | $a_0$, | $b_1$, | ?, | ?, | ? | $e_8 :$ | ?, | ?, | ?, | $f_0$, | ? |
| $e_6 :$ | ?, | ?, | $d_5$, | ?, | ? | $e_9 :$ | $a_0$, | $b_1$, | ?, | ?, | ? |
| $e_8 :$ | ?, | $b_1$, | $d_5$, | $f_0$, | ? | $e_{51} :$ | ?, | ?, | $d_5$, | ?, | $h_0$ |
| $e_{42} :$ | $a_0$, | ?, | $d_5$, | ?, | ? | $e_{98} :$ | ?, | ?, | $d_5$, | ?, | $h_0$ |

Hence, with $s_2$ it is never possible to pass from $a_0$ to $d_5$, whereas it is allowed by $s_1$ (you begin with $e_2$ to go from $a_0$ to $b_1$, and then you use $e_8$ to go from $b_1$ to $d_5$). It means that, in the case of $s_1$, you are always able to find a "correlation path" from one descriptor of the description to another, i.e. examples explaining the relationship between the descriptors in the stereotype.

# 3 Experiments

This section presents experiments performed on artificial data sets. This is followed by an original comparison in a real data case using three well-known clusterers. Default clustering was implemented in a Java program called PRESS (*Programme de Reconstruction d'Ensembles de Stéréotypes Structurés*). All the experiments for k-modes, EM and Cobweb were performed using the Weka platform [Garner, 1995].

## 3.1 Validation on artificial data sets

These experiments use artificial data sets to validate the robustness of our algorithm. The first step is to give some contrasted descriptions of $\mathcal{D}$. Let us note $n_s$ the number of these descriptions. Next, these initial descriptions are duplicated $n_d$ times. Finally, missing data are artificially simulated by removing a percentage $p$ of descriptors at random from these $n_s \times n_d$ artificial examples. The evaluation is carried out by testing different clusterers on these data and comparing the discovered cluster representatives with the initial descriptions. We verify what we call *recovered descriptors*, i.e. the proportion of initial descriptors that are found. This paper presents the results obtained with $n_s = 5$ and $n_d = 50$ over $50$ runs. The number of examples is $250$ and the descriptions are built using a langage with $30$ binary attributes. The tabu-list length is equal to $10$ and $NStep$ to $100$. Note that the first group of experiments are placed in the Missing Completely At Random (MCAR) framework.

Fig. 2 shows firstly that the results of PRESS are very good with a robust learning process. The stereotypes discovered correspond very well to the original descriptions up to 75% of missing descriptors. In addition, this score remains good (nearly 50%) up to 90%. Whereas Cobweb seems stable relative to the increase in the number of missing values, the results of EM rapidly get worse above 80%. Those obtained using k-modes are the worst, although the number of expected classes has to be specified.

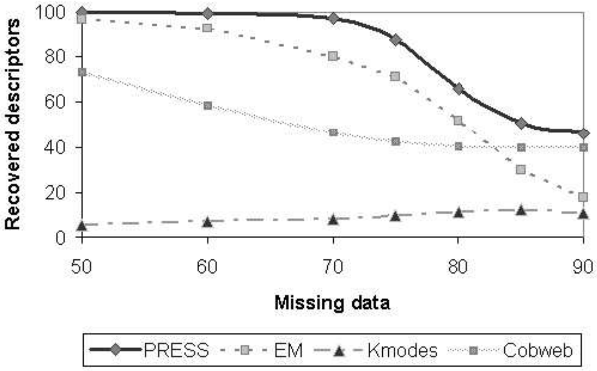

Figure 2: Proportion of recovered descriptors.

## 3.2 Studying social misrepresentation

The second part of the experiments deals with real data extracted from a newspaper called “Le Matin” at the end of the 19th century in France. The purpose is to automatically discover stereotype sets from events related to the political disorder in the first ten days of September 1893. The results of PRESS are compared to those of the three clusterers k-modes, EM and Cobweb. It should be pointed out that our interest focuses on the cluster descriptions, which we call *representatives* to avoid any ambiguity, rather than on the clusters themselves.

The articles linked to the chosen theme were gathered and represented using a language with 33 attributes. The terms of this language, i.e. attributes and associated values, were extracted manually. Most of the attributes are binary, 4 accept more than two values and 4 are ordinals. The number of extracted examples is 63 and the rate of missing data is nearly 87%, which is most unusual.

## 3.3 Evaluation of default clustering

In order to evaluate PRESS, a comparison was made with three classical clusterers: k-modes, EM and Cobweb. Hence, a non-probabilistic description of the clusters built by these algorithms was extracted using four techniques: (1) using the most frequent descriptors (mode approach); (2) the same as (1) but forbidding contradictory features between the examples and their representative; (3) dividing the descriptors between the different representatives; (4) the same as (3) but forbidding contradictory features. Two remarks need to be made. Firstly, the cluster descriptions resulting from k-modes correspond to technique (1). Nevertheless, we tried the other three techniques exhaustively. Secondly, representatives resulting from extraction techniques (3) and (4) entail *by construction* a redundancy rate of 0%. The comparison was made according to the following three points:

The first approach considers the contradictions between an example and its representative. The *example contradiction* is the percentage of examples containing at least one descriptor in contradiction with its covering representative. In addition, if you consider one of these contradictory examples, *average contradiction* is the percentage of descriptors in contradiction with its representative. This facet of conceptual clustering is very important, especially in the sparse data context.

| | k-Modes | | | | EM | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (1) | (2) |
| n | 6 | 6 | 6 | 6 | 2 | 2 |
| ex. contradiction | 27 | 0 | 27 | 0 | 48 | 0 |
| av. contradiction | 42 | 0 | 44 | 0 | 56 | 0 |
| $h_E$ | .89 | .60 | .74 | .50 | .85 | .66 |
| redundancy | 70 | 63 | 0 | 0 | 17 | 7 |
| cog. cohesion | × | × | × | × | × | × |

| | EM | | Cobweb | | | | PRESS |
|---|---|---|---|---|---|---|---|
| | (3) | (4) | (1) | (2) | (3) | (4) | |
| n | 2 | 2 | 2 | 2 | 2 | 2 | 6 |
| ex. cont. | 48 | 0 | 56 | 0 | 57 | 0 | 0 |
| av. cont. | 56 | 0 | 52 | 0 | 51 | 0 | 0 |
| $h_E$ | .83 | .65 | .82 | .56 | .68 | .46 | .79 |
| red. | 0 | 0 | 72 | 55 | 0 | 0 | 0 |
| cog. coh. | × | × | × | × | × | × | ✓ |

Figure 3: Comparative results on *Le Matin*.

Secondly, we check if the cognitive cohesion constraint (see 2.5) is verified. The rate of descriptor redundancy is also considered. These two notions are linked to the concept of stereotype and to the sparse data context.

Finally, we consider the degree of similarity between the examples and their covering representatives. This corresponds to the notion of compactness within clusters, but without penalizing the stereotypes with many descriptors. The function $h_E$ seems really adapted to give an account of representative relevance. In fact, we used a version of $h_E$ normalized between 0 and 1, by dividing by the total number of descriptors.

## 3.4 Results

Fig. 3 gives the results obtained from the articles published in Le Matin. Experiments for the k-modes algorithm were carried out with $N = 2 \ldots 8$ clusters, but only $N = 6$ results are presented in this comparison. The rows of the table show the number $n$ of extracted representatives, the two scores concerning contradiction, the result of $h_E$, the redundancy score and whether or not the cognitive cohesion constraint is verified. The columns represent each type of experiment (k-modes associated with techniques (1) to (4), EM and Cobweb as well, and finally our algorithm PRESS).

Let us begin by considering the contradiction scores. They highlight a principal result of default clustering: using PRESS, the percentage of examples having contradictory features with their representative is always equal to 0%. In contrast, the descriptions built using techniques (1) and (3) (whatever the clusterer used) possess at least one contradictory descriptor with 27% to 57% of the examples belonging to the cluster. Furthermore, around 50% of the descriptors of these examples are in contradiction with the covering description, and that can in no way be considered as a negligible noise. This is the reason why processes (1) and (3) must be avoided, especially in the sparse data context, when building such representatives from k-modes, EM or Cobweb clustering. Hence, we only consider techniques (2) and (4) in the following experiments.

Let us now study the results concerning clustering quality. This quality can be expressed thanks to the compactness function $h_E$, the redundancy rate and cognitive cohesion.

PRESS marked the best score (0.79) for cluster compactness with six stereotypes. That means a very good homogeneity between the stereotypes and the examples covered. It is perfectly consistent since our algorithm tries to maximize this function. The redundant descriptors rate is equal to $0\%$, according to the no-redundancy constraint. Furthermore, PRESS is the only algorithm that is able to verify cognitive cohesion. EM obtains the second best score and redundant descriptor rate remains acceptable. However, the number of expected classes must be given or guessed using a cross-validation technique, for instance. K-modes and Cobweb come third and fourth and also have to use an external mechanism to discover the final number of clusters.

Note that the stereotypes extracted using PRESS correspond to the political leanings of the newspaper. For instance, the main stereotype produces a radical, socialist politician, corrupted by foreign money and Freemasonry, etc. It corresponds partly to the difficulty in accepting the major changes proposed by the radical party and to the fear caused in France since 1880 by the theories of Karl Marx. We cannot explain here in more detail the semantics of discovered stereotypes, but these first results are really promising.

## 4 Conclusion

Conceptual clustering is seldom studied with such a high number of missing values. However, it is really important to be able to extract readable, understandable descriptions from such type of data in order to complete information, to classify new observations quickly and to make predictions. In this way, default clustering presented in this paper tries to provide an alternative to the usual clusterers. Moreover, based on local optimization techniques, it proposes a very general easy-to-extend framework for stereotype set discovering: new movements, constraints added relative to the problematic chosen, adapted control indexes, etc. The results obtained, on both artificial data sets and a real case extracted from newspaper articles, are really promising and should lead to other historical studies concerning social stereotypes.

For instance, we are currently applying these techniques to the study of *social representations*, a branch of social psychology introduced by S. Moscovici [Moscovici, 1961]. More precisely, this approach is really useful for *press content study* which up to now is done manually by experts. Hence, future work could be done on choosing key dates of the Dreyfus affair and automatically extracting stereotypical characters from different newspapers. These results will then be compared and contrasted with the work of sociologists and historians of this period.

## References

[Al-Sultan, 1995] K. Al-Sultan. A tabu search approach to the clustering problem. *Pattern Recognition*, 28(9):pp.1443–1451, 1995.

[Amossy and Herschberg Pierrot, 1997] R. Amossy and A. Herschberg Pierrot. *Stéréotypes et clichés: langues, discours, société*. Nathan Université, 1997.

[Figueroa *et al.*, 2003] A. Figueroa, J. Borneman, and T. Jiang. Clustering binary fingerprint vectors with missing values for DNA array data analysis. *IEEE Computer Society Bioinformatics Conference*, 2003.

[Fisher, 1987] D.H. Fisher. Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning*, (2):pp.139–172, 1987.

[Ganascia and Velcin, 2004] J.-G. Ganascia and J. Velcin. Clustering of conceptual graphs with sparse data. *Proceedings of the 12th International Conference on Conceptual Structures*, 2004.

[Garner, 1995] S.R. Garner. Weka: The waikato environment for knowledge analysis. *Proc. of the New Zealand Computer Science Research Students Conference*, pages pp.57–64, 1995.

[Gennari, 1990] J.H. Gennari. An experimental study of concept formation. 1990. Doctoral dissertation.

[Glover and Laguna, 1997] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, 1997.

[Lippman, 1922] W. Lippman. *Public Opinion*. Ed. MacMillan, NYC, 1922.

[Michalski, 1980] R.S. Michalski. Knowledge acquisition through conceptual clustering : A theorical framework and algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, (4):pp.219–243, 1980.

[Moscovici, 1961] S. Moscovici. *La psychanalyse : son image et son public*. PUF, Paris, 1961.

[Narboux, 2001] J.-P. Narboux. Ressemblances de famille, caractères, critères. pages pp.69–95. PUF, 2001.

[Newgard and Lewis, 2002] C.D. Newgard and R.J. Lewis. The Imputation of Missing Values in Complex Sampling Databases: An Innovative Approach. *Academic Emergency Medicine*, 9(5484), 2002.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, (13):pp.81–132, 1980.

[Rich, 1979] E. Rich. User Modeling via Stereotypes. *International Journal of Cognitive Science*, 3:pp.329–354, 1979.

[Rosch, 1975] E. Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, (104):pp.192–232, 1975.

[Rosch, 1978] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. NJ: Lawrence Erlbaum, Hillsdale, 1978.

[Sarkar and Leong, 2001] M. Sarkar and T.Y. Leong. Fuzzy k-means clustering with missing values. *Proc AMIA Symp.*, pages pp.588–92, 2001.

[Shawver, 2004] L. Shawver. Commentary on Wittgenstein's Philosophical Investigations, 2004. http://users.rcn.com/rathbone/lw65-69c.htm.

[Wittgenstein, 1953] L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, UK, 1953.