# Expert Constrained Clustering: a Symbolic Approach[⋆]

Fabrice Rossi[⋆⋆] and Frédérick Vautrain

LISE/CEREMADE (CNRS UMR 7534), Université Paris-IX/Dauphine,
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
e-mails: rossi@ceremade.dauphine.fr, vautrain@ceremade.dauphine.fr

**Abstract.** A new constrained model is discussed as a way of incorporating efficiently *a priori* expert knowledge into a clustering problem of a given individual set.
The first innovation is the combination of fusion constraints, which request some individuals to belong to one cluster, with exclusion constraints, which separate some individuals in different clusters. This situation implies to check the existence of a solution (ie if no pair of individuals are connected by fusion and exclusion constraints).
The second novelty is that the constraints are expressed in a symbolic language that allows compact description of group of individuals according to a given interpretation.
This paper studies the coherence of such constraints at individual and symbolic levels. A mathematical framework, close to the Symbolic Data Analysis[3], is built in order to define how a symbolic description space may be interpreted on a given individual set. A partial order on symbolic descriptions (which is an usual assumption of Artificial Intelligence), allows a symbolic analysis of the constraints. Our results provide an individual but also a symbolic clustering.

## 1 Introduction

In order to take into account prior expert knowledge, it is quite common to implement **constraints** in classification algorithms [4]. The general goal of classification is to find a satisfying partition of the population (set of individuals). Adding constraints allows to reduce the set of acceptable partitions. In this paper we consider only fusion and exclusion constraints. A fusion constraint implies to keep specified individuals into one cluster. For instance, we will not accept partitions in which individuals with a given property (e.g., small size) are classified into separated clusters. An exclusion constraint is exactly the opposite: we ask for specified individuals to remain in distinct clusters.

[⋆] Published in PKDD 2000.
Available at http://apiacoa.org/publications/2000/pkdd2000.pdf

[⋆⋆] Up to date contact informations for Fabrice Rossi are avaible at http://apiacoa.org/

It is possible to take into account fusion and exclusion constraints at the same time, which introduces a **coherency problem**: some sets of fusion and exclusion constraints cannot be satisfied together by any partition. For instance, if we ask for $x$ to belong to $y$'s cluster (fusion) and also for $x$ to be in a different cluster than $y$ (exclusion), it is obvious that no partition can satisfy both constraints. The coherency problem (solution existence) can be solved with a simple algorithm, as shown in [2]. The proposed algorithm works on a population and checks if a constraint set is coherent. When it is, the algorithm gives the smallest partition which satisfies the constraints (a partition $P$ is smaller than a partition $Q$ when each cluster of $P$ is included in a cluster of $Q$). One of the drawbacks of this algorithm is that it works directly at the individual level, which can be slow on big populations because the global analysis is needed to compute the smallest partition: it is not simple in general to work on subsets of the population and to merge the results. Moreover, working at the population level gives results that can be difficult to analyze.

In order to improve this algorithm, we propose to work at a *symbolic level*. We will describe individual subsets thanks to a symbolic approach that allows short representation of such subsets. For instance, a subset $C$ of individuals might be defined thanks to a conjunction of properties that all the individuals of $C$ satisfy, e.g., individuals whose weight belongs to a specified interval.

A mapping called ext (for *extension*) interprets each symbolic description on the set of individuals. It allows to build the actual individual subset associated to a description. Each constraint is defined as a pair of symbolic descriptions $(d, d')$. The interpretation mapping translates each pair into a classical constraint on the population, as follows:

1. if $(d, d')$ is a fusion constraint then, individuals which are described by $d$ and individuals which are described by $d'$ must belong to one cluster;
2. if $(d, d')$ is an exclusion constraint then, individuals which are described by $d$ and individuals which are described by $d'$ must belong to different clusters;

In this paper, we prove that under reasonable assumptions on the description space, the coherency problem can be studied directly at the symbolic level. The main advantage of this approach it to reduce the processing cost because each description can cover a lot of individuals. Moreover, when the constraints are coherent, the proposed algorithm gives a symbolic description of the smallest partition. Another interesting point is that the interpretation mapping can be changed without modifying the coherency result, as long as it stays in a broad class of acceptable mappings.

The rest of the paper is organized as follows: we start by defining the proposed mathematical framework. Then we present three important results. First we provide a mixed approach that allows to study symbolic constraints in relation to their interpretation on a given set of individuals. Then, we give our pure symbolic coherency results.

Due to size constraints, proof are omitted and can be found in [6].

# 2 Mathematical framework

## 2.1 Description space and context definition

The considered constrained classification problem involves:

1. the **description space** (called $\mathcal{D}$) corresponds to high level (symbolic) description of individuals (or groups of individuals);
2. the **population** (called $\Omega$) is a set of individuals that can be described thanks to the description space;
3. the **interpretation mapping** (called ext) is a map from $\mathcal{D}$ to $\mathcal{P}(\Omega)$ (the set of all subsets of $\Omega$). It transforms a symbolic description into a set of individuals that are correctly described by this description. For a description $a$, $\mathrm{ext}(a)$ is called the **extension** of $a$.
   The pair $\mathcal{G} = (\Omega, \mathrm{ext})$ is called a **context** for $\mathcal{D}$.

In general, the description space $\mathcal{D}$ is fixed, but the population or the interpretation mapping can change. For instance, if $\mathcal{D}$ uses very high level description such as "*tall*", the exact interpretation of "*tall*" might depend on the chosen semantic (e.g., if "*tall*" is applied to human beings, it might describe men higher than 1m80 at current time, or men higher than 1m70 several centuries ago). If the population is stored in a database the extension mapping can be considered as a two step process: first we translate the symbolic description into a SQL SELECT query and then we let the database compute the corresponding result which is the extension of the description.

In general, it is possible to provide an order denoted $\leq$ for $\mathcal{D}$ (see for instance [5] for examples of such symbolic orders). In our framework, we need such an order to be able to provide pure symbolic analysis (see section 4). The only technical assumption needed on the ordered set $(\mathcal{D}, \leq)$ is that any totally ordered subset of $\mathcal{D}$ must be bounded below.

The intuitive meaning of the symbolic order is linked to description precision. More precisely, $a \leq b$ means that $a$ is less general than $b$. In other words, $a$ is a particular case of $b$, for instance "*blue*" is less general than "*red or blue*". The interpretation mapping **must** translate this meaning into an equivalent property on the population. Technically, we have:

**Definition 1** *An* ***ordered description space*** *is a pair $(\mathcal{D}, \leq)$, where $\mathcal{D}$ is a description space and $\leq$ an order on $\mathcal{D}$. If $\Omega$ is a population and* ext *is a mapping from $\mathcal{D}$ to $\mathcal{P}(\Omega)$, the context $\mathcal{G} = (\Omega, \mathrm{ext})$ is* ***consistent*** *with $(\mathcal{D}, \leq)$ if and only if*

1. ext *is an increasing mapping from $(\mathcal{D}, \leq)$ to $(\mathcal{P}(\Omega), \subset)$;*
2. *for any $a$ and $b$ in $\mathcal{D}$:*

$$\min(a, b) = \emptyset \Rightarrow \mathrm{ext}(a) \cap \mathrm{ext}(b) = \emptyset, \tag{1}$$

   *where* $\min(a, b) = \{c \in \mathcal{D} \mid c \leq a \text{ and } c \leq b\}$.

Keeping in mind that $u \leq v$ means $u$ is less general than $v$, $\min(a,b) = \emptyset$ means that no description can be more precise than $a$ and $b$ at the same time. Therefore, if there are some individuals in $\mathrm{ext}(a) \cap \mathrm{ext}(b)$, this means that there is no way to describe those individuals more precisely than with $a$ or $b$. This is a symptom of *inadequacy* of the description space to the context and we assume therefore that this condition cannot happen.

**Example** Let's consider a population $\Omega$ a subset of $\textsc{Color} \times \textsc{Weight}(g) \times \textsc{Height}(cm)$, where $\textsc{Color} = \{Yellow, Blue, Red\}$, $\textsc{Weight}(g) = [1, 100]$ and $\textsc{Height}(cm) = [1, 25]$. Each individual $\omega$ (a spare part) is defined as a triplet $\omega = (\omega_{\mathrm{C}}, \omega_{\mathrm{W}}, \omega_{\mathrm{H}})$.

Let $\mathcal{D} = \mathcal{P}^*(\textsc{Color}) \times \mathcal{I}^*(\textsc{Weight}) \times \textsc{Height}$ a set of symbolic descriptions where $\mathcal{P}^*(\textsc{Color})$ is the set of the subsets (parts) of $\textsc{Color}$ minus the set $\varnothing$, $\mathcal{I}^*(\textsc{Weight})$ is the set of the subsets (intervals) of $\textsc{Weight}$ minus the set $\varnothing$ and $\textsc{Height}= \{Small, Tall\}$.

We denote a symbolic description $d = (d_{\mathrm{C}}, d_{\mathrm{W}}, d_{\mathrm{H}})$ and the symbolic order on $\mathcal{D}$ is defined as follows:

$$d \preceq d' \Leftrightarrow d_{\mathrm{C}} \subseteq d'_{\mathrm{C}}, d_{\mathrm{W}} \subseteq d'_{\mathrm{W}}, d_{\mathrm{H}} = d'_{\mathrm{H}}$$

An interpretation mapping (ext) can be defined for instance as follows:

$$\begin{cases} \mathrm{ext}\left((d_{\mathrm{C}}, d_{\mathrm{W}}, Small)\right) = \{\omega \in \Omega \mid \omega_{\mathrm{C}} \in d_{\mathrm{C}}\ ,\ \omega_{\mathrm{W}} \in d_{\mathrm{W}}\ ,\ 1 \leq \omega_{\mathrm{H}} < 15\} \\ \mathrm{ext}\left(d_{\mathrm{C}}, d_{\mathrm{W}}, Tall\right) = \{\omega \in \Omega \mid \omega_{\mathrm{C}} \in d_{\mathrm{C}}\ ,\ \omega_{\mathrm{W}} \in d_{\mathrm{W}}\ ,\ 15 \leq \omega_{\mathrm{H}} \leq 25\} \end{cases}$$

### 2.2 Constraints

**Definition 2** *A* ***constraint set*** *for a description space $\mathcal{D}$ is a pair of subsets of $\mathcal{D}^2$, $(F, E)$. The first subset $F$ represents* ***fusion*** *constraints and the second subset $E$ represents* ***exclusion*** *constraints.*

### 2.3 Constrained binary relation

**Definition 3** *Let $\mathcal{D}$ be a description space and $(\Omega, \mathrm{ext})$ be a context for $\mathcal{D}$. Let $(F, E)$ be a constraint set for $\mathcal{D}$. A binary relation $r$ on $\Omega$ is* ***compatible*** *with $(F, E)$ if and only if:*

$$\forall (a,b) \in F, \forall x \in \mathrm{ext}(a),\ \forall y \in \mathrm{ext}(b),\ r(x,y) \tag{2}$$

$$\forall (c,d) \in E, \forall x \in \mathrm{ext}(c),\ \forall y \in \mathrm{ext}(d),\ \neg r(x,y) \tag{3}$$

Our goal is to study constrained classification: in general, we will focus on equivalence binary relations.

### 2.4 Notations

Let $R$ be a subset of $T^2$, an arbitrary set. $R$ is the graph of a binary relation on $T$ and we can define the following sets:

- ${}^tR$ is the transposed (or dual) relation defined by: $(x, y) \in {}^tR \Leftrightarrow (y, x) \in R$;
- $R^s$ is the symmetric closure of $R$, defined by: $R^s = R \cup {}^tR$;
- if $S$ is a subset of $T^2$, $RS$, the product (or composition) of $R$ and $S$, is defined as follows: $RS = \{(a, b) \in T^2 \mid \exists c \in T \text{ such that } (a, c) \in R \text{ and } (c, b) \in S\}$
- $R^k$ is the $k$-th power of $R$, defined as follows: $R^0 = I_T$ ($I_T$ is the diagonal of $T$, i.e., $I_T = \{(x, x) \mid x \in T\}$) and $R^k = \{(x, y) \in T^2 \mid \exists z \in T \text{ so that } (x, z) \in R^{k-1} \text{ and } (z, y) \in R\}$, for $k > 0$. Thanks to the previous definition, this can be rewritten in $R^k = R^{k-1}R$;
- $R^+$ is the transitive closure of $R$, defined by: $R^+ = \bigcup_{k \geq 1} R^k$;
- $R^*$ is the reflexive and transitive closure of $R$, defined by: $R^* = R^+ \cup I_T$;
- $\mathcal{S}(R)$ is the support of $R$, defined by: $\mathcal{S}(R) = \{x \in T \mid \exists y \in T \text{ so that } (x, y) \in R \text{ or } (y, x) \in R\}$, which can be simplified into $\mathcal{S}(R) = \{x \in T \mid \exists y \in T \text{ so that } (x, y) \in R^s\}$.

## 3 Population based analysis

The purpose of this section is to analyze the coherency problem at symbolic level, but taking into account the interpretation and the population.

### 3.1 Fusion constraint closure

To state our results, we need first to introduce some technical definitions:

**Definition 4** *Let $R$ and $S$ be two binary relations on the same set $T$. $R$ is said to be **stable with respect to** $S$ if and only if for all $a$, $b$, $c$ and $d$ in $T$, we have:*

$$(a, b) \in R,\ (b, c) \in S \text{ and } (c, d) \in R \Rightarrow (a, d) \in R, \tag{4}$$

*i.e., $RSR \subset R$.*

In informal terms, this means that $S$ does not introduce short-circuits in $R$.

**Definition 5** *Let $R$ and $S$ be two binary relations on the same set $T$. We call $R_S$ the closure of $R$ by $S$, which is defined as the smallest binary relation that contains $R$ and that is stable with respect to $S$. Let us define $R_S^0 = R$ and for $k > 0$, $R_S^k = R_S^{k-1}SR$. We have the following properties:*

1. *$R_S = \bigcup_{k \geq 0} R_S^k$;*
2. *is $R$ is a symmetric relation on $T$ and is $S$ is a symmetric relation, then $R_S$ is a symmetric relation;*
3. *is $R$ is a transitive relation on $T$, then $R_S$ is a transitive relation.*

**Definition 6** *Let $F$ be a binary relation on $\mathcal{D}$ a description space and let $\mathcal{G} = (\Omega, \mathrm{ext})$ be a context for $\mathcal{D}$.*

*We denote $F_\mathcal{G} = \{(a,b) \in F \mid \mathrm{ext}(a) \neq \emptyset, \mathrm{ext}(b) \neq \emptyset\}$ ($E_\mathcal{G}$ is defined in a similar way).*

*We denote $\widetilde{F_\mathcal{G}}$ the closure of $((F_\mathcal{G})^s)^+$ by $S$, where $S$ is defined by:*

$$S = \{(a,b) \in \mathcal{D} \mid \mathrm{ext}(a) \cap \mathrm{ext}(b) \neq \emptyset\} \tag{5}$$

*$\widetilde{F_\mathcal{G}}$ is a symmetric and transitive relation. Moreover, for each $a \in \mathcal{S}(F_\mathcal{G})$, $(a,a) \in \widetilde{F_\mathcal{G}}$, which means that $\widetilde{F_\mathcal{G}}$ is an equivalence relation on $\mathcal{S}(F_\mathcal{G})$.*

This closure is an useful tool for classification study: if we have $(a,b) \in F$, where $F$ is the fusion part of a constraint set, this means that elements in $\mathrm{ext}(a)$ and $\mathrm{ext}(b)$ must be related by compatible binary relations. If $(c,d) \in F$, the same property is true for $\mathrm{ext}(c)$ and $\mathrm{ext}(d)$. Let us assume now that $\mathrm{ext}(b) \cap \mathrm{ext}(c) \neq \emptyset$, and let $y$ be an element of this intersection. Let now $x$ be an element of $\mathrm{ext}(a)$ and $z$ an element of $\mathrm{ext}(d)$. If $r$ is a compatible binary relation on $\Omega$, we have $r(x,y)$ and $r(y,z)$. Therefore, if $r$ is transitive (this is the case for classification, where $r$ is an equivalence relation), we have $r(x,z)$. Therefore, $r$ is also compatible with $F \cup \{(a,d)\}$ and we have, by construction, $(a,d) \in \widetilde{F_\mathcal{G}}$.

### 3.2 Constraint set coherency on a population

**Definition 7** *Let $(F,E)$ be a constraint set on $\mathcal{D}$ a description space and let $\mathcal{G} = (\Omega, \mathrm{ext})$ be a context for $\mathcal{D}$. We say that $(F,E)$ is* ***coherent on*** *$\mathcal{G}$, and we note $F \lhd_\mathcal{G} E$, if and only if the following conditions are satisfied:*

1. *for all $(a,b) \in E$, $\mathrm{ext}(a) \cap \mathrm{ext}(b) = \emptyset$;*
2. *for all $(a,b) \in F$ and $(c,d) \in E$, $\mathrm{ext}(a) \cap \mathrm{ext}(c) = \emptyset$ or $\mathrm{ext}(b) \cap \mathrm{ext}(d) = \emptyset$.*

We can now give our first result:

**Theorem 1** *Let $(F,E)$ be a constraint set on $\mathcal{D}$ a description space and let $\mathcal{G} = (\Omega, \mathrm{ext})$ be a context for $\mathcal{D}$. The following properties are equivalent:*

1. *there is a binary equivalence relation on $\Omega$ compatible with $(F,E)$;*
2. *$\widetilde{F_\mathcal{G}} \lhd_\mathcal{G} E$.*

*Moreover, the smallest equivalence relation on $\Omega$ compatible with $(F,E)$ can be defined as follows: $r(x,y)$ if and only if there are $(a,b) \in \widetilde{F_\mathcal{G}}$ with $x \in \mathrm{ext}(a)$ and $y \in \mathrm{ext}(b)$.*

### 3.3 Discussion

The practical implications of theorem 1 are important. It builds an equivalence relation on $\mathcal{S}(F_\mathcal{G})$, a subset of the description space. This equivalence relation is used for two purposes:

1. it gives a simple criterion for the existence of a solution to the constrained clustering problem;
2. the smallest clustering is the "image" of the symbolic equivalence relation by ext.

In a database system, the advantages of this approach are obvious: rather than working on the global population, we simply have to compute intersection of extensions. As in general, computing an extension is simply a SQL SELECT query, it is straightforward to calculate the intersection query. Therefore we never need to extract the full population from the database, but only to make some queries to first build $\mathcal{S}(F_\mathcal{G})$ and then to check if $\widetilde{F_\mathcal{G}} \lhd_\mathcal{G} E$.

Whereas this approach gives results easier to interpret than the pure population based approach given in [2], it is still based on the population and results depend on the context $\mathcal{G}$.

# 4 Pure symbolic analysis

In this section, we use the symbolic order so as to provide a pure symbolic answer to the coherency problem: given a constraint set $(F, E)$ on an ordered description space $(\mathcal{D}, \leq)$, can we prove that $(F, E)$ is coherent enough so that for any consistent context $\mathcal{G}$, there will exist a compatible equivalence relation (i.e., a solution to the constrained clustering problem).

## 4.1 Description set based closure

**Definition 8** *Let $F$ be a binary relation on $(\mathcal{D}, \leq)$ an ordered description space. We call $\widetilde{F}$ the closure of $(F^s)^+$ by $S$, where $S$ is defined by:*

$$S = \{(a, b) \in \mathcal{D} \mid \min(a, b) \neq \emptyset\} \tag{6}$$

*$\widetilde{F}$ is a symmetric and transitive relation. Moreover, for each $a \in \mathcal{S}(F)$, $(a, a) \in \widetilde{F}$, which implies that $\widetilde{F}$ is an equivalence relation on $\mathcal{S}(F)$.*

## 4.2 Constraint set coherency

**Definition 9** *Let $(F, E)$ be a constraint set on $(\mathcal{D}, \leq)$ an ordered description space. We say that $(F, E)$ is **coherent on** $\mathcal{D}$, and we note $F \lhd E$, if and only if the following conditions are satisfied:*

1. *for all $(a, b) \in E$, $\min(a, b) = \emptyset$;*
2. *for all $(a, b) \in F$ and $(c, d) \in E$, $\min(a, c) = \emptyset$ or $\min(b, d) = \emptyset$.*

**Theorem 2** *Let $(F, E)$ be a constraint set on $(\mathcal{D}, \leq)$ an ordered description space such that any totally ordered subset of $\mathcal{D}$ is bounded below. The following properties are equivalent:*

1. *$\widetilde{F} \lhd E$*
2. *for each consistent context $\mathcal{G} = (\Omega, \mathrm{ext})$, there exists a equivalence relation $r$ on $\Omega$ compatible with $(F, E)$*

### 4.3 Discussion

The practical implications of theorem 2 are quite important:

1. the computation is done at a pure symbolic level, which reduces in general the cost compared to previous approaches (if the population is stored in a database, the pure symbolic analysis makes no access to this database);
2. if obtained, the coherency result applies to *any* consistent context, which avoids to assume that the studied population is exhaustive or fixed (for instance). Moreover, this allows to change the interpretation, if needed;
3. the equivalence relation built on $\mathcal{S}(F)$ does not give directly (through the interpretation mapping) the smallest compatible equivalence relation on a given context, but only one among many possible equivalence relations. Some clusters are constructed by the closure calculation, based on stability through min. On a particular context, it might happen that $\min(a, b) \neq \emptyset$ but that $\operatorname{ext}(a) \cap \operatorname{ext}(b) = \emptyset$. If the description space is well suited to describe the population, a possible interpretation is to say that the population is not exhaustive. For instance, if we consider "*Red or Blue*" and "*Red or Yellow*" descriptions, the description space will in general contain a "*Red*" description (this is for instance the case in the example defined in section 2.1). It might happen that the population does not contain a "*Red*" individual. At the symbolic level, such individual potentially exists, and **must** be taken into account. The equivalence relation induced by $\widetilde{F}$ take them into account and is therefore bigger than the smallest that might available. The symbolic approach shows therefore relationship that are hidden on a particular context.

## 5 Conclusion

In this paper, we have introduced a new approach for studying symbolic constrained classification. This approach allows to work both on a population and at a pure symbolic level. The symbolic tools give efficient algorithms (especially when the population is big) as well as easy result analysis.

The proposed symbolic approach has been implemented in the SODAS framework [1] and is currently being benchmarked against the population based approach.

## References

1. Hans-Hermann Bock and Edwin Diday, editors. *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, 2000.
2. Roland De Guio, Thierry Erbeja, and Vincent Laget. A clustering approach for GT family formation problems. In *1st international conference on Engineering Design and Automation*, pages 18–21, March 1997.

3. Edwin Diday. L'analyse des données symboliques : un cadre théorique et des outils. Technical Report 9821, LISE/CEREMADE (CNRS UMR 7534), Université Paris-IX/Dauphine, Mars 1998.
4. A. D. Gordon. A survey of constrained classification. *Computational Statistics & Data Analysis*, 21:17–29, 1996.
5. Amedeo Napoli. Une introduction aux logiques de descriptions. Technical Report 3314, Projet SYCO, INRIA Lorraine, 1997.
6. Fabrice Rossi and Frédérick Vautrain. Constrained classification. Technical report, LISE/CEREMADE (CNRS UMR 7534), Université Paris-IX/Dauphine, April 2000.