Homogeneity and Stability in Conceptual Analysis

Paula Brito[1] and Géraldine Polaillon[2]

[1] Faculdade de Economia & LIAAD/INESC-Porto L.A., Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal mpbrito@fep.up.pt
[2] SUPELEC Science des Systèmes (E3S) - Département Informatique
Plateau de Moulon, 3 rue Joliot Curie, 91192 Gif-sur-Yvette cedex, France
geraldine.polaillon@supelec.fr

Abstract. This work comes within the field of data analysis using Galois lattices. We consider ordinal, numerical single or interval data as well as data that consist on frequency/probability distributions on a finite set of categories. Data are represented and dealt with on a common framework, by defining a generalization operator that determines intents by intervals. In the case of distribution data, the obtained concepts are more homogeneous and more easily interpretable than those obtained by using the maximum and minimum operators previously proposed. The number of obtained concepts being often rather large, and to limit the influence of atypical elements, we propose to identify stable concepts using interval distances in a cross validation-like approach.

1 Introduction

This work concerns multivariate data analysis using Galois concept lattices. Let $E = \{\omega_1, \ldots, \omega_n\}$ be the set of elements to be analyzed, described by p variables $Y_1, \ldots, Y_p$. In this paper we consider the specific case where the variables Y_j are numerical (real or interval-valued), ordinal and modal. Modal variables allow associating with each element of E a probability/frequency distribution on an underlying finite set of categories (see [9]).

The use of Galois lattices in Data Analysis was first introduced by Barbut and Monjardet, in the seventies of last century [2] and then further developed and largely spread out by the work of R. Wille and B. Ganter (see, e.g., [6]). Let $(A, \leq_1)$ and $(B, \leq_2)$ be two ordered sets. A Galois connection is a pair (f, g), where f is a mapping $f : A \to B$, g is a mapping $g : B \to A$, such that f and g are antitone, and both $h = gof$ and $h' = fog$ are extensive; h and h' are then closure operators. The mapping f defines the intent of a set $S \subseteq E$, and the mapping g that allows obtaining the extent in E associated with a set of attributes $T \subseteq O$, where O is the set of the considered (binary) attributes. The couple (f, g) then constitutes a Galois connection between $(P(E), \subseteq)$ and $(P(O), \subseteq)$. A concept is defined as a couple (S, T) where $S \subseteq E, T \subseteq O, S = g(T)$ and $T = f(S)$, i.e., we have $h(S) = S$; S is the *extent* of the concept and T its *intent*. This approach has been applied to non-binary variables, but in this case data are generally submitted to a previous "binarization", by performing a binary coding of the

data array; for numerical or ordinal variables Y, attributes of the form "$Y \leq x$," for each observed value x, are considered.

In [3] this approach has been extended by defining directly the intent of a set of elements; which has allowed obtaining, for each variable type (classical or otherwise) appropriate couples of mappings (f, g) forming a Galois connection. This has the advantage of allowing analyzing the data directly as it is presented, without imposing any sort of binary pre-coding, which may, and generally does, drastically increase the size of the data array to be analyzed. Galois lattices where intents are obtained by union and by intersection are obtained. This approach has been further extended to modal variables (see [4]). The case of ordinal variables has been dealt with in [11], using an approach similar to that of [4] for modal variables.

Ganter and Kuznetsov [5] proposed a general construction, called pattern structures, which allows for arbitrary descriptions with a semilattice operation on them; since union and intersection of intervals define semilattices, they make respective pattern structures. An application on gene expression data is presented in [7].

Here, we consider a common framework for numerical (real or interval-valued), ordinal and modal variables, by defining a generalization operator that determines the intent in the form of vectors of intervals. For ordinal and modal (i.e., distribution-valued) variables the obtained concepts are more homogeneous and therefore easier to interpret than those obtained by applying the minimum and maximum operators, as previously proposed. In the next sections, we detail how generalization of a set of elements is performed for each variable type.

The number of obtained concepts being often rather large, we propose to identify stable concepts (see also [8] and [12]), using distances designed for interval data. The criteria is that the intent of a concept should not be too different from those obtained by sequentially removing one element of the extent at a time - which would reveal that this particular element is provoking a drastic change in the concepts' intent. Should it occur, the concept would be considered to be non-stable.

In the case of multi-valued data, other approaches of lattice reduction, directly applied to the concept lattice, have been proposed in [1] and [10]. These two approaches rely on the same idea of merging together similar attribute values (in respect to a given threshold), and thereby reducing the number of concepts.

The remainder of the paper is organized as follows. Section 2 describes the generalization procedure for real and interval-valued variables, which is extended in Section 3 to modal variables. In Section 4 a common generalization approach by vectors of intervals is presented. In Section 5 the problem of concept stability is considered, and a method using interval distances is proposed, which allows addressing the question of lattice reduction. Section 6 concludes the paper, opening paths for future research.

2 Real and interval-valued variables

Let $E = \{\omega_1, ..., \omega_n\}$ be the set of n elements or objects to be analyzed, and $Y_1, \ldots, Y_p$ real or interval-valued variables such that $Y_j(\omega_i) = [l_{ij}, u_{ij}]$. We shall consider real-valued variables as a special case of interval-valued ones; it is therefore equivalent to write $Y_j(\omega_i) = x$ or $Y_j(\omega_i) = [x, x]$.

Let $A = \{\omega_1, \ldots, \omega_h\} \subseteq E$. Generalization by union is defined (see [3]) by the mapping $f : P(E) \to I^p$ where I is the set of intervals of $\mathbb{R}$ endowed with the inclusion order, such that $f(A) = (I_1, \ldots, I_p)$, with $I_j = [Min\{l_{ij}\}, Max\{u_{ij}\}]$, $\omega_i \in A$, $j = 1, \ldots, p$, i.e., for each $j = 1, \ldots, p, I_j$ is the minimum interval (for the inclusion order) that covers all values taken by the elements of A for variable Y_j. Let $g : I^p \to P(E)$ be the mapping defined as $g((I_1, \ldots, I_p)) =$
$= \{\omega_i \in E : Y_j(\omega_i) \subseteq I_j, j = 1, \ldots, p\}$, i.e., the set of elements of E taking values within I_j, for $j = 1, \ldots, p$. The couple (f, g) is a Galois connection.

Likewise, we may generalise by intersection defining f and g as follows: $f^* : P(E) \to I^p$, $f(A) = (I_1, \ldots, I_p)$, with $I_j = [Max\{l_{ij}\}, Min\{u_{ij}\}]$ if $Max\{l_{ij}\} \leq Min\{u_{ij}\}, \omega_i \in A$, $I_j = \oslash$ otherwise (i.e., the largest interval contained in all intervals taken by the elements of A for variable Y_j, which may be empty), for $j = 1, \ldots, p$, and $g^* : I^p \to P(E)$ with $g^*((I_1, \ldots, I_p)) = \{\omega_i \in E : Y_j(\omega_i) \supseteq I_j, j = 1, \ldots, p\}$ (the set of elements of E taking interval-values that contain I_j,) for $j = 1, \ldots, p$. The couple (f^*, g^*) forms also a Galois connection.

Example 1:
Consider three persons, Ann, Bob and Charles characterized by two variables, age and amount of time (in minutes) necessary to go to work (which varies from day to day, and is therefore represented by an interval-valued variable), as presented in Table 1.

	Age	Time
Ann	25	[15, 20]
Bob	32	[25, 30]
Charles	40	[10, 20]

Table 1. Age and amount of time (in minutes) necessary to go to work for three persons.

Let $A = \{$Bob,Charles$\}$. Generalization by the union leads to
$f(A) = ([32, 40], [10, 30])$, describing people who are between 32 and 40 years old and take 10 to 30 minutes to go to work; in this dataset people meeting this description are given by $g(f(A)) = g(([32, 40], [10, 30]))$, i.e., $\{$Bob, Charles$\} = A$. Here, $(\{$Bob, Charles$\}, ([32, 40], [10, 30]))$ is a concept.

3 Modal variables

Two Galois connections may also be defined for the case of modal variables (see [4]). Let $Y_1, \ldots, Y_p$ be p modal variables, $O_j = \{m_{j1}, \ldots, m_{jk_j}\}$ the set of k_j possible categories of variable Y_j, M_j the set of distributions defined on O_j, for $j = 1, \ldots, p$, and $M = M_1 \times \ldots \times M_p$. For variable Y_j and element $\omega_i \in E$, $Y_j(\omega_i) = \left\{m_{j1}(p_{j1}^{\omega_i}), \ldots, m_{jk_j}(p_{jk_j}^{\omega_i})\right\}$, where $p_{jk_\ell}^{\omega_i}$ is the probability/frequency associated with category $m_{j\ell}$ ($\ell = 1, \ldots, k_j$) of variable Y_j, and element ω_i. Let $A = \{\omega_1, \ldots, \omega_h\} \subseteq E$.

To generalise by the maximum we take, for each category $m_{j\ell}$, the maximum of its probabilities/frequencies in A. Let $f : P(E) \to M$, such that $f(A) = (d_1, \ldots, d_p)$, with $d_j = \{m_{j1}(t_{j1}), \ldots, m_{jk_j}(t_{jk_j})\}$, where $t_{j\ell} = Max\{p_{j\ell}^{\omega_i}, \omega_i \in A\}, \ell = 1, \ldots, k_j$. The intent of a set $A \subseteq E$ is then to be interpreted as "objects with *at most* $t_{j\ell}$ cases presenting category $m_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p$". The couple (f, g) with $g : M \to P(E)$ defined as, for $d_j = \{m_{j1}(p_{j1}), \ldots, m_{jk_j}(p_{jk_j})\}$, $g((d_1, \ldots, d_p)) = \left\{\omega_i \in E : p_{j\ell}^{\omega_i} \leq p_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p\right\}$, forms a Galois connection.

Similarly, we may generalise by the minimum taking for each category the minimum of its probabilities/frequencies. Let $f^* : P(E) \to M$, $f^*(A) = (d_1, \ldots, d_p)$, with $d_j = \{m_{j1}(v_{j1}), \ldots, m_{jk_j}(v_{jk_j})\}$, where $v_{j\ell} = Min\{p_{j\ell}^{\omega_i}, \omega_i \in A\}, \ell = 1, \ldots, k_j$. The intent of a set $A \subseteq E$ is now interpreted as "objects with *at least* $v_{j\ell}$ cases presenting category $m_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p$".
The couple (f^*, g^*) with $g^* : M \to P(E)$ such that, for $d_j = \{m_{j1}(p_{j1}), \ldots, m_{jk_j}(p_{jk_j})\}$, $g^*((d_1, \ldots, d_p)) = \left\{\omega_i \in E : p_{j\ell}^{\omega_i} \geq p_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p\right\}$ forms likewise a Galois connection.

Example 2:
Consider four groups of students for each of which a categorical mark is given, according to the following scale: a: mark < 10, b: mark between 10 and 15, c: mark > 15 as summarized in Table 2.

	Mark
Group 1	$< 10(0.2), [10-15]\,(0.6), > 15(0.2)$
Group 2	$< 10(0.3), [10-15]\,(0.3), > 15(0.4)$
Group 3	$< 10(0.1), [10-15]\,(0.6), > 15(0.3)$
Group 4	$< 10(0.3), [10-15]\,(0.6), > 15(0.1)$
Group 5	$< 10(0.5), [10-15]\,(0.3), > 15(0.2)$

Table 2. Frequency distributions of the students marks, in 3 categories, for 5 groups.

The intent, obtained by the maximum operator, of the set formed by groups 1 and 2, is $\{a(0.3), b(0.6), c(0.4)\}$ and is interpreted as "students' groups with *at*

most 30% of marks a, *at most* 60% of marks b and *at most* 40% of marks c". The corresponding extent comprehends groups 1, 2, 3 and 4. If, alternatively, we determine the intent of the same set by the minimum operator, we obtain $\{a(0.2), b(0.3), c(0.2)\}$, to be read as "students' groups with *at least* 20% of marks a, *at least* 30% of marks b and *at least* 20% of marks c", whose extent is formed by groups 1, 2 and 5.

4 A common approach: generalization by intervals

We now present a unique framework allowing to perform generalization for numerical (real or interval-valued) variables, ordinal variables and modal variables, based on generalization by intervals.

For numerical (real or interval-valued) data, we are in the above mentioned case of generalization by taking the union.

For modal variables, it amounts to consider, for each category, an interval corresponding to the range of its probability/frequency. In fact, it has often been observed that generalization either by the maximum or by the minimum, as defined in Section 3, may quickly lead to over-generalization. As a consequence, $f(A)$, $A \subseteq E$, is not very informative.

Let $M_j^I = \{m_{j\ell}(I_{j\ell}), \ell = 1, \ldots, k_j\}, m_{j\ell} \in O_j, I_{j\ell} \subseteq [0,1]$ and $M^I = M_1^I \times \ldots \times M_p^I$. Generalization is now defined as

$$f^I : P(E) \to M^I$$
$$f^I(A) = (d_1, \ldots, d_p)$$
$$\text{with } d_j = \{m_{j1}(I_{j1}), \ldots, m_{jk_j}(I_{jk_j})\},$$

where $I_{j\ell} = \left[Min\{p_{j\ell}^{\omega_i}\}, Max\{p_{j\ell}^{\omega_i}\}\right], \omega_i \in A, \ell = 1, \ldots, k_j$, $j = 1, \ldots, p$ and

$$g^I : M^I \to E$$
$$g((d_1, \ldots, d_p)) = \left\{\omega_i \in E : p_{j\ell}^{\omega_i} \in I_{j\ell}, \ell = 1, \ldots, k_j, j = 1, \ldots, p\right\}$$

The so-defined couple of mappings (f^I, g^I) forms a new Galois connection.

On the data of Example 2, generalization by intervals of groups 1 and 2 provides the intent $\{a\,[0.2, 0.3]\,, b\,[0.3, 0.6]\,, c\,[0.2, 0.4]\}$, to be read as "students' groups having between 20% and 30% cases of mark a, between 30% and 60% cases of mark b and between 20% and 40% cases of mark c" and whose extent now only contains groups 1 and 2.

The case of ordinal variables has been addressed in [11], performing generalization either using the maximum or the minimum. To allow for more flexibility, the author proposes to choose the operator individually for each variable. Nevertheless, one of these generalization operators must be chosen in each case, and over-generalization is not prevented. Our proposal for this type of variables, is to generalise a set $A \subseteq E$ considering, no longer a minimum or a maximum, but rather an interval of ordinal values.

Example 3:
Consider the classifications given by four cinema critics while evaluating three movies, Movie 1, Movie 2 and Movie 3 as given in Table 3.

	Movie 1	Movie 2	Movie 3
Critic 1	5	5	4
Critic 2	5	4	4
Critic 3	1	2	2
Critic 4	2	1	1

Table 3. Classifications given by four critics to three movies.

The intent obtained by using the maximum operator of the group formed by critics 1 and 2 is $(5, 5, 4)$, to be interpreted as "critics giving *at most* mark 5 to Movie 1, *at most* mark 5 to Movie 2 and *at most* mark 4 to Movie 3" - which is obviously too general and would cover almost everyone; in this dataset the corresponding extent contains critics 1, 2, 3 and 4. Therefore, the class formed by critics 1 and 2, who present a similar behavior, does not correspond to a concept. The intent obtained by using the minimum operator of the group formed by critics 3 and 4 is $(1, 1, 1)$, to be read "critics giving *at least* mark 1 to Movie 1, *at least* mark 1 to Movie 2 and *at least* mark 1 to Movie 3" - which would cover every critic; its extent in this dataset consists again of critics 1, 2, 3 and 4. Here again, the class formed by critics 3 and 4, who give quite similar marks, does not correspond to a concept. If we now perform generalization by interval-vectors of the group formed by critics 1 and 2, we obtain the intent $([5, 5], [4, 5], [4, 4])$; likewise for the group formed by critics 3 and 4, we have $([1, 2], [1, 2], [1, 2])$; in the first case we are clearly referring to critics giving high marks while in the second case we describe critics giving low marks to all movies. The corresponding extents no longer contain other critics, presenting a rather different profile from those considered each time. Furthermore, both ({Critic 1, Critic 2}, $([5, 5], [4, 5], [4, 4])$ and ({Critic 3, Critic 4}, $([1, 2], [1, 2], [1, 2])$ are concepts. When determining concepts, according to the minimum or the maximum operators, e.g. in a clustering context, there is therefore a risk of forming heterogeneous clusters, since over-generalization may lead to a too large extent. By taking interval-vectors of observed values, the over-generalization problem is avoided.
To conclude this section, we now present a more general example, with variables of the different considered types.

Example 4:
Consider the data in Table 4, where 4 persons are described by their age, a real-valued variable, time (in minutes) they take to go to work, an interval-valued variable, the means of transportation used, a modal variable, and their classifications given to three newspapers, A, B and C (ordinal variable).

	Age	Time	Transport	A	B	C
Albert	25	$[15, 20]$	car (0.2) bus (0.8))	4	2	5
Bellinda	40	$[25, 30]$	car (0.7), bus (0.2), train (0.1))	2	4	3
Christine	32	$[10, 15]$	car (0.2), bus (0.7), train (0.1))	5	1	4
David	58	$[30, 45]$	car (0.9), bus (0.1))	2	4	1

Table 4. Age, time taken to go to work (in minutes), means of transportation used and classifications given to newspapers A, B and C for four persons.

The intent of $A = \{\text{Albert, Christine}\}$ is
$V = ([25, 32]\ ,\ [10, 20]\ ,\ ([0.2, 0.2], [0.7, 0.8], [0.0, 0.1])\ ,\ [4, 5], [1, 2], [4, 5])$ and (A, V) is a concept.

5 Stability

Concepts are theoretically very interesting, and do provide rich information on the values shared by subsets of elements of the set under study. However, the number of concepts of a data array is often rather large, even for relatively low cardinals of the sets of elements and variables. This fact makes the analysis and interpretation of results a bit delicate. It is often to be noticed that when analyzing the concepts generated by numerical or modal variables, groups of concepts appear which are quite similar. This may be due to noise or minor differences, generally not pertinent. The idea is therefore to extract only those concepts which are representative of these groups of similar concepts, so as to obtain a more concise representation with significantly homogeneous concepts.

Several solutions may be pointed out for this objective. We will focus on the notion of stability, as introduced in [8] and [12], which evaluates the amount of information of the intent that depends on specific objects of the concept's extent. Formally, the stability of a concept is defined as the probability of keeping its intent unchanged while deleting arbitrarily chosen objects of its extent.

When analyzing data described by numerical (real or interval-valued), ordinal or modal variables, and generalizing using interval-vectors (as described in the previous sections), we shall apply a similar approach to each formed concept, but introducing a distance measure. The objective being to retain the homogeneous concepts, it is wished to avoid that a single element of the concepts' extent produces an important increase in the intent's intervals' ranges.

To identify the stable concepts, a threshold α depending on the maximum distance is defined (so as no to be dependent from the variables' scales). A concept is said to be "stable" if the distance between the intent obtained by removing one element of the extent at a time, and its original intent, is not above the given threshold. This is in fact a cross-validation-like approach, in that one element of the extent is removed at a time, and the resulting intent is compared with the original one.

When data have an interval form, interval distances should be used. Different measures are available in the literature; we will focus on three interval distance measures: the *Hausdorff distance*, the *interval Euclidean distance* and the *interval City-Block distance*.

Let $I_i = [l_i, u_i]$ and $I_h = [l_h, u_h]$ be two intervals we wish to compare. The *Hausdorff distance* d_H, the *interval Euclidean distance* d_2 and the *interval City-Block distance* d_1 between I_i and I_h are respectively

$$d_H(I_i, I_h) = Max\left\{\left\{\left|l_i - l_h\right|, \left|u_i - u_h\right|\right\}\right.$$
$$d_2(I_i, I_h) = \sqrt{(l_i - l_h)^2 + (u_i - u_h)^2}$$
$$d_1(I_i, I_h) = \left|l_i - l_h\right| + \left|u_i - u_h\right|.$$

The *Hausdorff distance* between two sets is the maximum distance of a set to the nearest point in the other set, i.e., two sets are close in terms of the Hausdorff distance if every point of either set is close to some point of the other set. *Interval Euclidean* and *City-Block* distances are just the counterparts of the corresponding distances for real values; if we embed the interval set in $\mathbb{R}^2$, where one dimension is used for the lower and the other for the upper bound of the intervals, then these distances are just the *Euclidean* and *City-Block* distances between the corresponding points in the two-dimensional space.

Let $C = (A, D)$ be a concept, where $A = \{\omega_1, \ldots, \omega_h\} \subseteq E$ is its extent and $D = (I_1, \ldots, I_p)$ is its intent, $D = f(A)$. The considered criterion is then the distance Δ between D et D^{-i} where D^{-i} is the intent of A without ω_i, $D^{-i} = f(A \setminus \{\omega_i\}), i = 1, \ldots, h$, defined by: $\Delta = Max\{\delta(D, D^{-i}), \omega_i \in A\}$, δ measuring the dissimilarity between interval-vectors.

Let d be the distance (according to the chosen measure) between the intervals corresponding to variable Y_j in a concept's intent. Two options may then be foreseen, whether it is wished to consider the maximal or the average distance on the intervals defining the intents:

1. $\delta_{Max}(D, D^{-i}) = Max\{d(I_j, I_j^{-i})\}$, j indexing the variable set $Y_j, j = 1, \ldots, p$ in the case of numerical and ordinal variables, and the global category set $O = O_1 \cup \ldots \cup O_p$ in the case of p modal variables;
2. $\delta_{Mean}(D, D^{-i}) = Mean\{d(I_j, I_j^{-i})\}$, j as in 1.

A concept $C = (A, D)$ is then considered to be stable if $\Delta \leq \alpha$. This approach allows keeping only the stable, and therefore more representative, concepts, avoiding the effect of outlier observations.

6 Illustrative application

Consider again classifications given by cinema critics evaluating three movies, Movie 1, Movie 2 and Movie 3 where $Y_j(Critic_i)$ is the mark given by Critic i to Movie j, $i = 1, \ldots, 5; j = 1, 2, 3$, as given in Table 5.

Tables 6 and 7 list the concepts obtained when the Minimum and the Maximum generalization operators are used, respectively.

	Movie 1	Movie 2	Movie 3
Critic 1	3	2	3
Critic 2	1	1	2
Critic 3	5	5	1
Critic 4	4	3	2
Critic 5	2	4	5

Table 5. Classifications given by five critics to three movies.

	Intent		
Extent	Movie 1	Movie 2	Movie 3
$\{1\}$	≥ 3	≥ 2	≥ 3
$\{3\}$	≥ 5	≥ 5	≥ 1
$\{4\}$	≥ 4	≥ 3	≥ 2
$\{5\}$	≥ 2	≥ 4	≥ 5
$\{1,4\}$	≥ 3	≥ 2	≥ 2
$\{1,5\}$	≥ 2	≥ 2	≥ 3
$\{3,4\}$	≥ 4	≥ 3	≥ 1
$\{3,5\}$	≥ 2	≥ 4	≥ 1
$\{1,3,4\}$	≥ 3	≥ 2	≥ 1
$\{1,4,5\}$	≥ 2	≥ 2	≥ 2
$\{3,4,5\}$	≥ 2	≥ 3	≥ 1
$\{1,2,4,5\}$	≥ 1	≥ 1	≥ 2
$\{1,3,4,5\}$	≥ 2	≥ 2	≥ 1
$\{1,2,3,4,5\}$	≥ 1	≥ 1	≥ 1

Table 6. Concepts of the Minimum lattice corresponding to the data in Table 5.

	Intent		
Extent	Movie 1	Movie 2	Movie 3
$\{2\}$	≤ 1	≤ 1	≤ 2
$\{3\}$	≤ 5	≤ 5	≤ 1
$\{1,2\}$	≤ 3	≤ 2	≤ 3
$\{2,4\}$	≤ 4	≤ 3	≤ 2
$\{2,5\}$	≤ 2	≤ 4	≤ 5
$\{1,2,4\}$	≤ 4	≤ 3	≤ 3
$\{1,2,5\}$	≤ 3	≤ 4	≤ 5
$\{2,3,4\}$	≤ 5	≤ 5	≤ 2
$\{1,2,3,4\}$	≤ 5	≤ 5	≤ 3
$\{1,2,4,5\}$	≤ 4	≤ 4	≤ 5
$\{1,2,3,4,5\}$	≤ 5	≤ 5	≤ 5

Table 7. Concepts of the Maximum lattice corresponding to the data in Table 5.

The concepts (except for the empty extent one) obtained from this data table, using generalization by intervals, i.e., for $A \subseteq E, f(A) = (I_1, I_2, I_3)$, with $I_j = [Min\{Y_j(Critic_i)\}, Max\{Y_j(Critic_i)\}]$, $Critic_i \in A, j = 1, 2, 3$, are listed in Table 8.

	Intent		
Extent	Movie 1	Movie 2	Movie 3
$\{1\}$	$[3,3]$	$[2,2]$	$[3,3]$
$\{2\}$	$[1,1]$	$[1,1]$	$[2,2]$
$\{3\}$	$[5,5]$	$[5,5]$	$[1,1]$
$\{4\}$	$[4,4]$	$[3,3]$	$[2,2]$
$\{5\}$	$[2,2]$	$[4,4]$	$[5,5]$
$\{1,2\}$	$[1,3]$	$[1,2]$	$[2,3]$
$\{1,4\}$	$[3,4]$	$[2,3]$	$[2,3]$
$\{1,5\}$	$[2,3]$	$[2,4]$	$[3,5]$
$\{2,4\}$	$[1,4]$	$[1,3]$	$[2,2]$
$\{2,5\}$	$[1,2]$	$[1,4]$	$[2,5]$
$\{3,4\}$	$[4,5]$	$[3,5]$	$[1,2]$
$\{3,5\}$	$[2,5]$	$[4,5]$	$[1,5]$
$\{4,5\}$	$[2,4]$	$[3,4]$	$[2,5]$
$\{1,2,4\}$	$[1,4]$	$[1,3]$	$[2,3]$
$\{1,2,5\}$	$[1,3]$	$[1,3]$	$[2,5]$
$\{1,3,4\}$	$[3,5]$	$[2,5]$	$[1,3]$
$\{1,4,5\}$	$[2,4]$	$[2,4]$	$[2,5]$
$\{2,3,4\}$	$[1,5]$	$[1,5]$	$[1,2]$
$\{3,4,5\}$	$[2,5]$	$[3,5]$	$[1,5]$
$\{1,2,3,4\}$	$[1,5]$	$[1,5]$	$[1,3]$
$\{1,2,4,5\}$	$[1,4]$	$[1,4]$	$[2,5]$
$\{1,3,4,5\}$	$[2,5]$	$[2,5]$	$[1,5]$
$\{1,2,3,4,5\}$	$[1,5]$	$[1,5]$	$[1,5]$

Table 8. Concepts of the interval lattice for the data in Table 5.

We notice that all the concepts obtained using the Minimum or the Maximum operator are concepts for the interval generalization, although with a different meaning, given the different intent mapping. As discussed before, even in this small example it may be observed that concepts obtained using the Minimum or the Maximum operator often present a rather general intent, thus leading to over-generalization in the concept formation. Consider, for instance, the concept ($\{1\}$, (Movie $1 \geq 3$, Movie $2 \geq 2$, Movie $3 \geq 3$)) in Table 6, it indicates that Critic 1 gives high marks to each movie, which is not really the case, whereas the concept ($\{1\}$, (Movie $1 \in [3,3]$, Movie $2 \in [2,2]$, Movie $3 \in [3,3]$)) in Table 8 gives a much more accurate description of the concepts's extent. Also, concept ($\{3\}$, (Movie $1 \leq 5$, Movie $2 \leq 5$, Movie $3 \leq 1$)) in Table 7 describes Critic 3

as giving any marks to Movies 1 and 2, and low marks to Movie 3; using interval generalization we learn that the marks given by Critic 3 to Movies 1 and 2 are the highest and non other. Consider now concept ($\{3,4\}$, (Movie $1 \geq 4$, Movie $2 \geq 3$, Movie $3 \geq 1$)) in Table 6: the intent reports any mark for Movie 3 (in particular, high marks are possible); if we use interval generalization instead we obtain the concept ($\{3,4\}$, (Movie $1 \in [4,5]$, Movie $2 \in [3,5]$, Movie $3 \in [1,2]$ which more accurately describes the observed situation.

We now compare the concepts retained as stable with each of the three distances, using both δ_{Max} and δ_{Mean}, and a threshold value of 1 and 2. The identified stable concepts in each case, represented by the corresponding extent, are listed in Table 9.

Distance	Criterion	Threshold	Stable concepts (extent)
d_H	Max	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{1,4\},\{1,5\},\{3,4\},$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
	Mean	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\},$ $\{1,2,4\},\{1,2,5\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},$ $\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{3,4\},\{4,5\}$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},\{2,3,4\},\{3,4,5\}$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
d_2	Max	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{1,4\},\{1,5\},\{3,4\},$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
	Mean	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\},$ $\{1,2,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},$ $\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{3,4\},\{4,5\}$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},\{2,3,4\},\{3,4,5\},$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
d_1	Max	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{1,4\},\{1,5\},\{3,4\},$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
	Mean	1	$\{1\},\{2\},\{3\},\{4\},\{5\},\{1,4\},$ $\{1,2,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$
		2	$\{1\},\{2\},\{3\},\{4\},\{5\},$ $\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{3,4\},\{4,5\},$ $\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,4,5\},\{2,3,4\},\{3,4,5\},$ $\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{1,2,3,4,5\}$

Table 9. Stable concepts for different distances, criteria and threshold values.

As it may be seen from Table 9, for all distances and both criteria, a demanding threshold identifies a small number of stable concepts, therefore leading to an important reduction in the number of retained concepts; if we use a more liberal threshold, a larger number of concepts are retained as stable, as was to be expected. The maximum criterion is naturally more strict than the mean, which retains more concepts as stable, for all distances and both threshold values. Finally, in this example, no important difference appears between the results obtained for the different distance measures.

7 Conclusion

A common generalization procedure, for numerical, ordinal and modal variables, which uses a representation based on interval-vectors is presented. This allows defining more homogeneous concepts, than generalization operators that use the maximum and/or the minimum. The proposed approach for ordinal variables allows addressing recommendation systems, analyzing preference data tables. It would also be interesting to explore how the proposed generalization operator behaves in a supervised learning context.

The number of obtained concepts being often rather large, a method for identifying stable concepts is proposed, using a cross-validation-like approach. This allows avoiding the effect of atypical elements in the concepts' formation. Naturally, the value of the used threshold has an important influence in the rate of concept reduction. The next step will be to explore this methodology for larger data tables, so as to have a more accurate evaluation of its efficiency in concept reduction. Another issue interesting to investigate is the comparaison of the list of concepts with those obtained with a subset of the given variables. This then leads to the problem of variable selection in the context of Galois lattices construction and analysis. As concerns applications, we are particularly interested in analyzing real preference data, for application in recommendation systems.

References

[1] Z. Assaghir, M. Kaytoue, N. Messai and A. Napoli (2009). On the mining of numerical data with Formal Concept Analysis and similarity. In *Proc. Société Francophone de Classification*, pp. 121-124.

[2] Barbut, M. and B. Monjardet (1970). *Ordre et Classification, Algèbre et Combinatoire, Tomes I et II.* Paris: Hachette.

[3] Brito, P. (1994). Order structure of symbolic assertion objects. *IEEE Transactions on Knowledge and Data Engineering 6*(5), 830–835.

[4] Brito, P. and G. Polaillon (2005). Structuring probabilistic data by Galois lattices. *Math. & Sci. Hum. / Mathematics and Social Sciences 169*(1), 77–104.

[5] Ganter, B. and S.O. Kuznetsov (2001). Pattern structures and their projections. In: G. Stumme and H. Delugach (Eds.), *Proc. 9th Int. Conf. on Conceptual Structures, ICCS'01*, Lecture Notes in Artificial Intelligence, vol. 2120, pp. 129-142.

[6] Ganter, B. and R. Wille (1999). *Formal Concept Analysis, Mathematical Foundations*. Berlin: Springer.

[7] Kaytoue, M., S.O. Kuznetsov, A. Napoli and S. Duplessis (2011). Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*, Volume 181, Issue 10, 1989–2001.

[8] Kuznetsov, S. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence 49*(1-4), 101–115.

[9] Noirhomme-Fraiture, M. and P. Brito (2011). Far beyond the classical data models: Symbolic Data Analysis. *Statistical Analysis and Data Mining 4*(2), 157–170.

[10] Pernelle, N., M.-C. Rousset, and V. Ventos (2001). Automatic construction and refinement of a class hierarchy over multi-valued data. In L. De Raedt and A. Siebes (Eds.), *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Computer Science, pp. 386–398.

[11] Pfaltz, J. (2007). Representing numeric values in concept lattices. In J. Diatta, P. Eklund and M. Liquiere (Eds.), *Proc. Fifth International Conference on Concept Lattices and Their Applications*, pp. 260–269.

[12] Roth, C., S. Obiedkov and D. Kourie (2008). On succint representation of knowledge community taxonomies with Formal Concept Analysis. *International Journal of Foundations of Computer Science 19*(2), 383–404.