Characterisation of gradual itemsets through "especially if" clauses based on mathematical morphology tools

Amal Oudni[1,2] **Marie-Jeanne Lesot**[1,2] **Maria Rifqi**[1,2,3]

[1]Université Pierre et Marie Curie-Paris6, UMR 7606, LIP6, F-75005, Paris, France
[2]CNRS, UMR 7606, LIP6 F-75005, Paris, France
[3]Université Panthéon-Assas - Paris 2 F-75005, Paris, France

Abstract

Gradual itemsets of the form "*the more/less A, the more/less B*" summarise data through the description of their internal tendencies, identified as correlation between attribute values. This paper proposes to characterise gradual itemsets, enriching them with an additional clause introduced by the linguistic expression "especially if": they are of the form "*the more/less A, the more/less B, especially if* $J \in R$", where J is a set of attributes occurring in $A \cup B$ and R is a set of intervals defined for each attribute in J. The method proposed to automatically extract characterised gradual itemsets is based on appropriate mathematical morphology tools. The paper illustrates the relevance of the proposed approach on a real data set.

1. Introduction

Gradual itemsets provide information summarising data sets in the linguistic form "*the more/less A, the more/less B*", e.g. illustrated by a sentence such as "the closer the wall, the harder the brakes are applied". Initially introduced in the fuzzy implication formalism [1, 2, 3], they have then been interpreted as expressing constraints on the attribute covariations. Several interpretations of the constraints have been proposed, as regression [4], correlation of induced order [5, 6] or identification of compatible object subsets [7, 8]. Each interpretation leads to the definition of a support and to methods for the identification of the itemsets that are frequent according to the considered support definition.

In the case of categorical or fuzzy data, it has been proposed to enrich gradual itemsets by so-called strengthening clauses, linguistically introduced by the expression "all the more" [9]: strengthened gradual itemsets are of the form "*the more/less A, the more/less B, all the more C*", where C is a set of categorical or fuzzy modalities associated to data attributes. It can be illustrated by the example "the closer the wall, the harder the brakes are applied, all the more the higher the speed" where *high* is a fuzzy modality associated to the speed attribute. The additional clause defines a restriction of the data such that the gradual itemset is better satisfied on the data subset than on the whole data set.

This paper proposes another enrichment, in the case of numerical data. The proposed characterisation clause, linguistically introduced by the expression "especially if", takes the form of attributes occurring in the considered itemset, associated with intervals. It can be illustrated by a sentence as "the closer the wall, the harder the brakes are applied, especially if the distance to the wall $\in [0, 50]$m", or more generally "*the more/less A, the more/less B, especially if* $J \in R$", where J is a set of attributes occurring in $A \cup B$ and R a set of intervals defined for each attribute in J.

The main difference between strengthening and characterising clauses is the fact that the former consider predefined restrictions, defined by the, possibly fuzzy, presence of specific attribute values. The characterisation proposed in this paper applies to numerical data and extracts both attributes and appropriate intervals to define the data restriction. It can be considered that such a set of pairs made of attributes and associated intervals define a new categorical feature, with value 1 for data whose attribute values belong to the considered intervals and 0 otherwise. However, the computational cost of applying the strengthening approach [9] to such augmented data would be too high; therefore an integrated method that directly looks for appropriate intervals is proposed. It relies on the use of mathematical morphological tools.

The paper is organised as follows: Section 2 recalls the formalism of gradual itemsets and presents related works. Section 3 describes the interpretation of gradual itemset characterisation and its formalisation. Section 4 presents the proposed methodology based on mathematical morphological tools in the base case and Section 5 the post-processing steps required for the general case. Section 6 illustrates the results obtained on real data.

2. Context and related works

2.1. Gradual itemsets

This section recalls the definitions of gradual items and itemsets [9, 8] as well as the support definition

based on compatible data subsets [8].

Gradual itemsets Throughout the paper $\mathcal{D}$ denotes the data set. A *gradual item* A^* is made of an attribute A and a variation $* \in \{\geq, \leq\}$, which represents a comparison operator. A *gradual itemset* is then defined as a set of gradual items $I = \{(A_j, *_j), j = 1..k\}$, interpreted as their conjunction. It is associated to its length, k, defined as the number of attributes it involves, and the pre-order $\preceq_I$ it induces, defined as

$$o \preceq_I o' \text{ iff } \forall\ j \in [1,k]\ A_j(o) *_j A_j(o')$$

where $A_j(o)$ represents the value of attribute A_j for object o.

Extraction by identification of compatible subsets In this paper, we consider the interpretation of the co-variation constraint by identification of compatible subsets [7, 8]: it consists in identifying subsets D of $\mathcal{D}$, called *paths*, that can be ordered so that all data pairs of D satisfy the pre-order induced by the considered itemset. More formally, for an itemset $I = \{(A_j, *_j), j = 1..k\}$, $D = \{o_1, ..., o_m\} \subseteq \mathcal{D}$ is a path if and only if there exists a permutation π such that

$$\forall l \in [1, m-1], o_{\pi_l} \preceq_I o_{\pi_{l+1}}$$

Such a path is called *complete* if no object can be added to it without violating the order constraint imposed by I. $\mathcal{L}(I)$ denotes the set of complete paths associated to I. The set of maximal complete paths, i.e. complete paths of maximal length, is denoted

$$\mathcal{L}^*(I) = \{D \in \mathcal{L}(I) / \forall D' \in \mathcal{L}(I)\ |D| \geq |D'|\}$$

The gradual support of I, $GS_{\mathcal{D}}(I)$, is then defined as the length of its maximal paths divided by the total number of objects.

$$GS_{\mathcal{D}}(I) = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}(I)} |D| \quad (1)$$

I is a *valid* itemset if $GS(I) \geq s$, where s is a user-set threshold. The GRITE algorithm [8] constitutes an efficient method to extract such valid gradual itemsets.

2.2. Strengthened gradual itemsets

Strengthened gradual itemsets are enriched itemsets, to which a clause linguistically introduced by the expression "all the more" is added [9]. They can be illustrated by the example "the closer the wall, the harder the brakes are applied, all the more the higher the speed". The strengthening clause consists of values of categorical attributes or fuzzy modalities of fuzzy attributes. The interpretation in terms of reinforced presence, proposed in [9], considers such enriched itemsets as itemsets that are better satisfied when the data set is restricted to the objects possessing, possibly in a fuzzy weighted sense, the values required by the strengthening clause.

Similarly, the characterisation proposed in this paper compares the validity of the itemset evaluated over the whole data set with the one measured on a restriction of the data. Yet the considered restriction requires that attribute values remain within given interval bounds, and is not limited to the presence of predefined modalities given as the data descriptors: the method automatically extracts the relevant intervals.

This objective and its assumptions induce a difference in the nature of the considered data: characterised gradual itemsets do not apply to categorical or fuzzy data, but to numerical data. It can be noted that fuzzy data can be processed by the proposed method, but that the difficulty comes from the interpretation of the obtained results: imposing that membership degrees belong to an identified interval does not seem to have a natural satisfying semantics.

2.3. Identification of interval of interest

The proposed characterisation by interval restriction also relates to works that aim at identifying intervals of interest, as occurs for mining of quantitative association rules and for fuzzy partition elicitation.

Quantitative association rules Quantitative association rules are an extension of classical association rules to numerical attributes [10, 11]: in this case indeed, an item cannot be defined as an attribute value, because the notion of occurrence frequency for a numerical value is not relevant. An item is defined as a couple made of an attribute with an interval, e.g. (age, [27, 38]). It is then possible to compute the proportion of data possessing an item to evaluate its support, and thus to apply classical itemset mining algorithms.

In order to identify such intervals of interest, some methods rely on an a priori discretisation of quantitative attributes, e.g. defined as equi-width or equi-depth intervals [10, 12, 11]. Intervals of interest are then identified with the Apriori algorithm, applied to extended data where binary features for each interval are added to indicate whether the numerical value of an attribute belongs to the corresponding interval.

Other methods extract single intervals at a time, within the rule generation phase. The evaluation of candidate intervals of interest depends on the quality of the rules they induce, e.g. measured by support, confidence or gain, often leading to computationally extensive algorithms. In order to limit the computational cost, some approaches rely on restricted rule schemes [13, 14], e.g. limiting the

number of numerical attributes in the premisse and the conclusion. Other methods exploit genetic algorithms [15, 16] to increase the efficiency of the candidate interval exploration while relaxing the rule form.

Fuzzy partition identification The identification of intervals of interest is also involved in the induction of fuzzy decision trees, where the decision taken at a given node depends on the interval to which the attribute value describing a data belongs, with a fuzzy weighting scheme [17]. It relies on a fuzzy discretisation of the attribute ranges and a selection based on criteria such as fuzzy entropies. The main difference with the methods proposed for quantitative association rules comes from the supervised learning framework decision trees belong to, and the exploitation of class information to determine candidate intervals.

The approach proposed by [17] uses mathematical morphology tools [18] to identify class homogeneous intervals, tolerating some noise in the intervals through the application of appropriate alternated filters.

The method we propose to identify intervals to characterise gradual itemsets falls within a supervised framework and is based on a transcription of the data that associates each observed value to a class, depending whether the corresponding object belongs to the itemset path. It is thus more similar to the issue of fuzzy partition identification than to that of quantitative association rules. We propose to also exploit mathematical morphology tools, as detailed in the following sections.

3. Formalization of characterized gradual itemsets

This section discusses the interpretation and principle of gradual itemset characterization, illustrating it on an example. It then presents the proposed formalization.

3.1. Illustrative example and interpretation

Figure 1 represents a data set described with two attributes, for which the gradual itemset $I = A^{\geq}B^{\geq}$ is supported by the path represented by $\bullet$ data. Its gradual support is $14/23 = 60\%$. Now it can be visually observed that the covariation between A and B especially holds in the center part of the graph, whereas more noisy data occur for low A values and high A values. Indeed, if the data are restricted to objects for which A takes values in the interval $[32; 53]$, graphically delimited by the vertical lines on Figure 1, the support of the itemset increases to $9/10 = 90\%$. This motivates the extraction of the characterized itemset $A^{\geq}B^{\geq}$*; especially if* $A \in [32; 53]$.

More generally, we propose to interpret the characterisation of gradual itemsets as an increased validity when the data are restricted to the objects satisfying the characterisation clause. Yet, in order to be informative, such a characterisation should not restrict the data too drastically: it is easy to achieve 100% support, for instance restricting the data to a couple of data points satisfying the order induced by the considered gradual itemset. Yet the derived characterization would be too specific and not valuable. In the previous example, restricting the data to the smaller interval $[32; 42]$ increases the support to 100%, but leads to a less general characterisation.

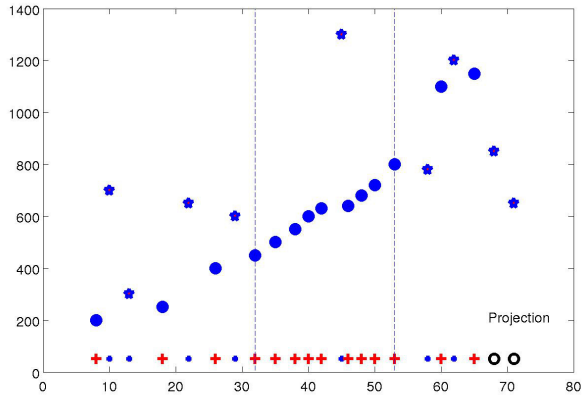

Figure 1: Example of gradual itemset characterisation, leading to the "the more A, the more B, especially if $A \in [32; 53]$.

The principle of characterised gradual itemsets is thus to find a trade-off between a high support and a high number of objects when restricting the data set to a subset.

3.2. Formalisation

Formally, the principle illustrated above can be presented as follows: for a gradual itemset I, e.g. extracted using the GRITE algorithm [8], a characterisation is denoted as "I*, especially if* $J \in R$", where J is a set of attributes occurring in I and R is an associated set of intervals. The region R induces a restriction $\mathcal{D}'$ of the data set $\mathcal{D}$, considering only the data satisfying the value constraint expressed by R.

The previous principle then consists in both maximising the support of the considered itemset I on the restricted data and the number of objects satisfying the constraints, i.e.

$$\begin{cases} \max\limits_{R} |\mathcal{D}'| \\ \max\limits_{R} GS_{\mathcal{D}'}(I) \end{cases} \tag{2}$$

A trade-off must be found between these two objectives that can be contradictory: indeed, an increase of the size of the subset $\mathcal{D}'$ can lead to the de-

crease of the proportion of objects compatible with the order induced by the considered itemset.

3.3. Proposed approach

We propose to decompose the task of identifying relevant attributes and their associated intervals of interest by successively considering each attribute occurring in the considered gradual itemset I and further by successively considering each path supporting I, e.g. available when I is extracted by GRITE [8]: the computation of the restricted gradual support $GS_{\mathcal{D}'}(I)$ can be based on the restriction of these paths. We thus propose to consider the effect of candidate restriction for each path, later combining them to select the optimal bounds.

Transcription Given a gradual itemset I, a maximal path D and the attribute A for which an interval of interest is looked for, and following the method existing for the elicitation of fuzzy partition [17], we encode the path information through a transcription process into a word composed of the symbols $\{+, -, \circ\}$. The ith character is obtained as the transcription of the ith object in the order induced by A, denoted x, as

- $x \to +$ iff $x \in D$
- $x \to -$ iff $(x \notin D) \wedge (A_{mD} \leq A(x) \leq A_{MD})$
- $x \to \circ$ otherwise

where A_{mD} and A_{MD} denote the minimal and maximal values of attribute A obtained for objects in D: $A_{mD} = \min_{x \in D} A(x)$ and $A_{MD} = \max_{x \in D} A(x)$. The $\circ$ symbol encodes data outside the boundaries of the processed path, it is necessary to handle the case of multiple maximal paths, as described in Section 5.1.

The lower part of Figure 1 indicates the transcription result for the illustrative example.

Word characteristics The objective formalised in Equation 2 can then be transposed to the path representation as a word: the restriction of the data set corresponds to a subpart of the word, and reciprocally, $|\mathcal{D}'|$ corresponds to the length of the subpart. The restricted support $GS_{\mathcal{D}'}(I)$ equals the proportion of $+$ symbols in this subpart.

Given a word v, we denote $l(v)$ its length and $NP(v)$ the number of $+$ it contains. The support definition of gradual itemsets is extended to words as $supp(v) = NP(v)/l(v)$.

The highest support is obtained when the considered subpart is a $+$ sequence containing no $-$ symbol, leading to $supp = 1$. In particular, the longest $+$ sequence observed in v, denoted $S(v)$, has for size $l(S(v))$ and its support is $supp(S(v)) = 1$.

The issue is then to extend such a sequence, incorporating some $-$ symbols, so as to increase the size of the restricted data set without deteriorating the proportion of $+$ in the considered subpart. It may for instance be the case that in v, two $+$ sequences s_1 and s_2, by definition shorter than $S(v)$, are only separated by a short $-$ sequence, denoted s_-. In this case, considering the word subpart made of the concatenation $s' = s_1 s_- s_2$ leads to a long sequence with still a high number of $+$. More formally, $l(s') = l(s_1) + l(s_-) + l(s_2)$ and $supp(s') = (l(s_1) + l(s_2))/l(s')$.

The trade-off between size and support then relates to the question whether one prefers to consider the data subset corresponding to $S(v)$, that maximises the support, or the one induced by s', which has a higher length at the expense of a lower support.

To that aim, we propose to exploit mathematical morphology tools as described in the next section.

4. Mathematical morphology tools for the identification of interval of interest

This section presents the mathematical morphology tools proposed to address the task presented in the previous section, as well as the analysis of their properties and relevance.

4.1. Principle

Mathematical morphology [18], denoted MM in the following, defines a set of tools for the identification of spatial structures as the shape and size of objects. It has been extensively used for image processing and functional analysis. One-dimensional MM [17], 1DMM, applies to words, obtained as data transcriptions on a set of symbols. The latter is $\{+, -, \circ\}$ in the case of interval of interest characterising gradual itemsets.

The aim is to smooth the considered words, ignoring isolated $-$ symbols that prevent from building large restricted data sets: indeed, it is then possible to increase the size of the considered subsequence, with a limited decrease of the proportion of $+$. As detailed below, such smoothing effects can be obtained when applying appropriate MM operators: the principle consists in applying an operator φ, leading to $v' = \varphi(v)$ in order to bridge gaps between $+$ sequences in v, identifying the longest $+$ sequence in v', $S(v')$, and evaluating the corresponding sequence in v, $S_v(v')$, with length $l(S_v(v'))$ and support $NP(S_v(v'))/l(S_v(v'))$.

Existing 1DMM tools have been proposed to get such smoothing effects in order to build fuzzy partitions in a supervised learning framework [17]. These operators transform a word defined on a binary symbol set, say $\{+, -\}$, to a ternary one $\{+, -, u\}$, where u denotes modified characters. The latter are interpreted as unstable regions, and thus fuzzy frontiers of the elicited fuzzy modalities.

In the case of characterising gradual subsets, the modified words are defined on the same symbol set, $\{+, -, \circ\}$, as the initial words. Moreover this set is

ternary from the beginning, because of the $\circ$ symbol that encodes data outside the boundaries of the processed path. The latter can be interpreted as bounds on the considered words and thus not modified by any considered operator. Another specificity of the characterising gradual subset issue is the absence of symmetry between the + symbol and the − symbol: the interest is entirely focused on + sequences, whereas in the case of fuzzy partition, + sequences and − sequences play equivalent roles.

4.2. Considered operators

This section describes the operators proposed to perform the desired mathematical morphology smoothing. They are transpositions of classic operators defined in image MM to the one-dimensional case.

Erosion Given a word defined on $\{+, -, \circ\}$, the erosion operator, denoted Er_1, decreases the size of + sequences replacing the outer + by −: for any $m \geq 0$

$$\begin{array}{ccccccccc} - & +^{m+2} & - & \longrightarrow & - & - & +^{m} & - & - \\ \circ & +^{m+1} & - & \longrightarrow & & \circ & +^{m} & - & - \\ - & +^{m+1} & \circ & \longrightarrow & - & - & +^{m} & \circ & \end{array}$$

The last two rows make explicit the specificity of the $\circ$ symbol.

For instance, for $v = - + + + - - - +$

$$Er_1(v) = - - + - - - - -$$

Er_n, where n is an integer parameter, denotes the combination of n successive erosions. It can be observed that the application of Er_n erases all + sequences of length lower than $2n$, as each of their elements is progressively replaced with −.

Dilatation Reciprocally the dilatation operator, denoted Di_1, decreases − sequences and expands + sequences: for any $m \geq 0$

$$\begin{array}{ccccccccc} + & -^{m+2} & + & \longrightarrow & + & + & +^{m} & + & + \\ \circ & -^{m+1} & + & \longrightarrow & & \circ & -^{m} & + & + \\ + & -^{m+1} & \circ & \longrightarrow & + & + & -^{m} & \circ & \end{array}$$

D_n is the combination of n successive dilatations. For instance, for the previous word v, $D_1(v)$ produces $Di_1(v) = + + + + + - + +$.

The application of Di_n erases all − sequences of length lower than $2n$.

Opening The opening operator is then defined, as in classical mathematical morphology, as $Op_n = Di_n \circ Er_n$. For example, with the word

$$\begin{aligned} v &= - - + + + + + - + + + - + \\ \text{one has } Op_1(v) &= - - + + + + + - + + + - - \\ \text{and } Op_2(v) &= - - + + + + + - - - - - - \end{aligned}$$

The effect of the erosion is to expand − sequences, the posterior dilatation makes it possible to reduce them again, except in the regions where the erosion step deleted all the + symbols. Indeed, in this case, there is no + symbol left to propagate. This means that, as compared to the initial word, the opening operator bridges the gap between − sequences separated by less than $2n$ + symbols.

Closure Reciprocally, the closure operator is defined as $Cl_n = Er_n \circ Di_n$. For example, starting from the word used in the previous example

$$\begin{aligned} \text{one has } Cl_1(v) &= - - + + + + + + + + + + - \\ \text{and } \quad Cl_2(v) &= - - + + + + + + + + + - - \end{aligned}$$

It bridges the gap between + sequences separated by less than $2n$ − symbols.

Alternated filter The alternated filter is the recursive combination of opening and closure operations:

$$\begin{aligned} n = 1 \quad & Filt_1 = Cl_1 \circ Op_1 \\ n > 1 \quad & Filt_n = Cl_n \circ Op_n \circ Filt_{n-1} \end{aligned}$$

For a given n, the combination $Cl_n \circ Op_n$ first deletes short + sequences, of length lower than $2n$, bridging the gap between − sequences. The remaining + sequence, that are thus of length greater than $2n+1$, can be grouped together by the closure operator if they are separated by less than $2n$ − symbols.

Moreover, the alternated filter is defined recursively, meaning that this behaviour is applied to the result of the previous filters, $Filt_{n-1} \circ Filt_{n-2} \circ \cdots \circ Filt_1$.

4.3. Properties

The analysis of the considered operators makes it possible to establish their properties, so as to examine the characteristics of the resulting word subpart extracted from the initial word.

Alternated filter trade-off behaviour It must first be underlined that the alternated filter is both more tolerant and more demanding when n increases: on one hand, it makes it possible to replace longer − sequences in + sequences, i.e. it is more tolerant to gaps within + blocks. On the other hand, to perform such modifications, it requires the longer − sequences to be surrounded by longer + sequences, i.e. it is more demanding for bridging gaps. This is the reason why it implements a trade-off between length and proportion of + symbols, providing an interesting tool for extracting intervals of interest.

This property is illustrated on Figure 2 which shows the support and the length of the extracted sequences from 10 random words and for filter sizes from 1 to 10: each line corresponds to one word, the connected points to the couples (support, length) after applying filters of increasing order, from $n = 1$ to $n = 10$. Two types of words can be observed: some of them are transformed to words with no + sequence, leading to null support and length. In these

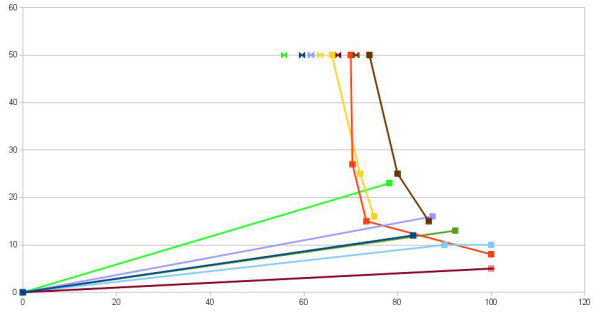

Figure 2: Support (x-axis) and length (y-axis) of the extracted sequences for 10 random words and alternated filter $Filt_1$ to $Filt_{10}$.

words, the + sequences are too short and bridged with low level filters. On the other hand, some other words show the trade-off between size and support: for low filter values, the + sequences are short and pure (i.e. high support, possibly 100%), when the filter order increases, more − symbols are erased, leading to longer extracted sequences but lower supports.

It can also be underlined, on the graph and more generally, that after applying the alternated filter $Filt_n$, either no + symbol at all occur in the obtained word, or the + sequence contains at least $2n+1$ + symbols. This partially gives a guarantee about the size of the extracted intervals when applying an alternated filter.

Alternated filter asymmetry From the observations given in the previous subsection, it can also be underlined that the combination of opening and closure leads to an interesting asymmetry of the $Filt_n$ filter. Indeed after the application of $Filt_n$,

- + sequences with length lower than $2n$ are replaced by − sequences
- − sequences with length lower than $2n$ and surrounded by + sequences of length greater than $2n+1$ are replaced by + sequences.

Defining short and long with respect to the threshold value $2n$, these properties show that short + sequences are unconditionally replaced by − sequences, whereas the replacement of short − sequences imposes conditions on the length of the surrounding + sequences (as previously defined in terms of tolerance and demand). This property is highly relevant in the context of gradual itemset characterisation, that focuses on + symbols, and is related to the requirement of not deteriorating the support value when lengthening the sequences.

Indeed, whereas a closure operator bridges the gap between + sequences independently of their lengths, possibly leading to long sequences with low support, the alternated filter only allows to bridge − sequences under the condition that they come along with even longer + sequence.

Some specific cases More precisely, in the worst case, i.e. the case with the lowest support, the combination $Cl_n \circ Op_n$ transforms in a + sequence the initial word $v = +^{2n+1} -^{2n} +^{2n+1}$. $S_v(v')$ has length $6n+2$ and support $(4n+2)/(6n+2)$, which is always greater than 0.66.

As a comparison, in the case with the lowest support, a closure operator transforms to a + sequence the initial word $+-^{2n}+$ of length $2n+2$ and support $2/(2n+2)$. The latter can be very low.

Examining the worst case of the alternated filter is more complex because of the recursive definition that can lead to a lower support than the value computed above. Indeed, as $Filt_n(v) = Cl_n \circ Op_n \circ Filt_{n-1}(v)$, it may be the case that the $+^{2n+1}$ sequence in the word $v = +^{2n+1} -^{2n} +^{2n+1}$ mentioned as worst case for $Cl_n \circ Op_n$ has been built by a bridging, or consolidating, effect of a previous filter, referring to less + symbols in the initial word.

Starting from low filter orders, this consolidation effect can be seen as follows: the consolidated sequence built by $Filt_1$ with the lowest number of + symbols is $u_1 = +^3 -^2 +^3$, of length 8. Thus, the consolidated sequence built by $Filt_2$ with the lowest number of + symbols is $u_2 = u_1 -^4 u_1$. More generally, denoting u_n the consolidated sequence built by $Filt_n$ with the lowest number of + symbols, one has the recursive relation $u_{n+1} = u_n -^{2n} u_n$. The sizes and supports of these sequences, respectively denoted C_n and S_n, verify

$$\begin{cases} C_1 = 8 \\ C_n = 2C_{n-1} + 2n \end{cases} \qquad \begin{cases} S_1 = 6 \\ S_n = 2S_{n-1} \end{cases}$$

5. Post-processing steps

The method described in the previous section presents the extraction of a relevant interval of interest for a given path. Now in the general case a gradual itemset is based on several complete paths, that can correspond to several characteristic intervals. This section describes the proposed aggregation operator to combine the results obtained from these paths.

5.1. Aggregation: processing multi-paths

A gradual itemset can rely on several paths, each one leading to a characteristic interval of interest; the latter must be aggregated to a single interval. We propose to perform an early aggregation, applied to the filtered words representing the paths: more precisely, the proposed aggregation function applies to words defined on $\{+, -, \circ\}$ having the same length, equal to the number of objects in the data set $|\mathcal{D}|$, obtained after filtering the transcription of paths as described in the previous section. It successively applies to each element of the sequence, and outputs a word defined on $\{+, \varnothing\}$. The $\varnothing$ symbol denotes values on which the itemset is not

characterised. This proposed aggregation function is defined as

$$\begin{array}{rcccc} Agg & : & \{+,-,\circ\}^2 & \rightarrow & \{+,\varnothing\} \\ & & (s_1,s_2) & \mapsto & s \end{array}$$

It is symmetrical and defined as follows, for all possible pairs of symbols

$$\begin{array}{r|l} s_1 & +++-\circ\circ \\ s_2 & +\circ---\circ \\ \hline Agg(s_1,s_2) & ++\varnothing\varnothing\varnothing\varnothing \end{array}$$

Values outside a path, denoted $\circ$, are neutral and do not influence the results; values that are excluded from a path are associated with $\varnothing$, i.e. excluded from the final result. This is compatible with the characterisation objective: only highly significant and representative elements are to be considered.

Figure 3 illustrates the aggregation step in the case of two paths, transcripted as v_1 and v_2, that present a high agreement when applying a filter of order $n = 1$. The aggregation thus leads to a large final + sequence S_c, whose bounds translated back to the attribute value domain define the interval of interest.

The paths considered for transcription and aggregation are the maximal paths in $\mathcal{L}^*(M)$. Indeed taking into account all complete paths $\mathcal{L}(M)$ could generate too many counter-examples and lead to a aggregated sequence only containing the $\varnothing$ symbol: in the transcription process, even if an object belongs to another path of the considered itemset, it is transcribed as $-$ if it does not belong to the processed path.

The characterisation interval is finally defined by its bounds, set as the minimal and maximal values of the considered attribute in the aggregated word.

5.2. Linguistic representation

The extracted intervals to characterise gradual itemsets are included in the corresponding linguistic summary, by default using a clause of the form "especially if", leading to the form "*the more/less A, the more/less B, especially if* $J \in R$".

When an interval J associated to an attribute A_J is very narrow, it seems to be relevant to replace the expression "especially if A_J is between $\min(J), \max(J)]$" by "especially if A_J equals J^*" where J^* is the central value of the interval.

The definition of the threshold defining whether an interval is narrow or not may be left to the user, to make the yielded summaries adapted to his/her needs and preferences.

6. Experimental study

This section describes the experiment carried out using the proposed characterisation method on a real data set. The analysis of the results is based on the comparison of the gradual support itemsets before and after characterisation.

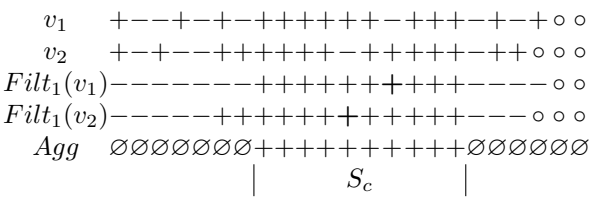

Figure 3: Aggregation obtained for two paths.

6.1. Considered data

We used a real data set called *weather* downloaded from the site http://www.meteo-paris.com/ile-de-france/station-meteo-paris/pro: these data come from the Parisian weather station of St-Germain-des-Prés. The data set contains 2133 meteorological observations realised during eight days (November 23^{rd} to 30^{th} 2012), described by 22 numerical attributes such as temperature (C), rain (mm), outside humidity (%), pressure (hPa), wind speed (km/hr) among others.

6.2. Comparison of $GS_{\mathcal{D}}$ and $GS_{\mathcal{D}'}$

We first extract gradual itemsets from the considered data, using the GRITE algorithm [8], setting the minimum gradual support $s = 20\%$ and defining the order induced by an attribute as a strict one: data with equal attribute values are not considered as supporting an itemset. This exclusion avoids the presence of characteristic intervals whose bounds are equal. We then apply the proposed characterisation methodology, setting the filter order $n = 4$.

The number of extracted itemsets before characterisation is 835; 509 are enriched by a characterisation clause, which corresponds to more than 60% of the extracted itemsets. Here are some examples of extracted gradual itemsets.

- The higher the temperature, the lower wind speed, especially if wind speed $\in [1, 10]$, $GS_{\mathcal{D}} = 36.3\%$, $GS_{\mathcal{D}'} = 78.6\%$
- The higher the pressure, the higher the temperature, especially if temperature $\in [13, 19.2]$, $GS_{\mathcal{D}} = 22\%$, $GS_{\mathcal{D}'} = 76\%$
- The lower the humidity, the lower the temperature, especially if temperature $\in [8.1, 12.2]$, $GS_{\mathcal{D}} = 22.8\%$, $GS_{\mathcal{D}'} = 70\%$

Figure 4 shows a comparison between the gradual support obtained before and after characterisation for each of the 509 extracted characterised gradual itemsets. All points being above the $y = x$ line, it shows that after characterisation, the gradual supports are higher than before, which confirms the increased validity of the gradual itemsets. The highest support obtained before characterisation is 42.4%; after characterisation the highest value is 78.6%, which is higher.

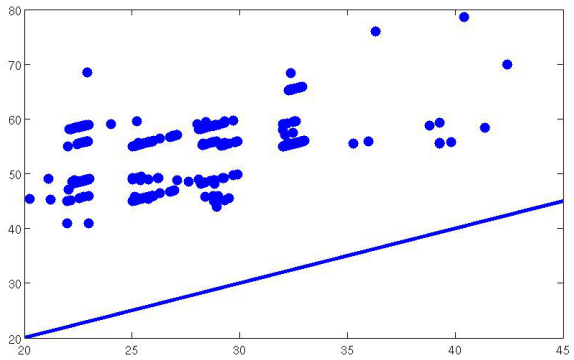

Figure 4: Comparison of the gradual support before (x-axis) and after (y-axis) characterisation, for each of the 509 extracted characterised gradual itemsets.

7. Conclusion and future work

In this paper we propose an approach to characterise gradual itemsets, linguistically expressed by the addition of a clause introduced by the expression "especially if", so as to extract more information summarising data sets. The extraction of characterised gradual itemsets relies on the identification of intervals of interest for the attributes occurring in the considered gradual itemset. The approach proposed to address this task is based on mathematical morphology tools to achieve a trade-off between an increased validity of the itemset when restricted to the interval and large intervals of interest, imposing meaningful characterisations

Future works aim at studying more formally the properties of the proposed mathematical morphology operators, in particular to establish guarantees on the minimum support and minimum length obtained for a given filter order. Perspectives also include applying characterisation using attributes that do not occur in the considered gradual itemset. This task raises issues regarding the time and memory consumption, so as to efficiently rule out non-relevant features as soon as they can be detected as such. Another perspective is to introduce density constraints so as to focus on relevant regions of the domain that are not too sparsely populated.

References

[1] S. Galichet, D. Dubois, H. Prade. Imprecise specification of illknown functions using gradual rules. *Int. Journal of Approximate Reasoning, 2004*. vol. 35, pp. 205–222.

[2] E. Hüllermeier. Implication-based fuzzy association rules. *Principles of Data Mining and Knowledge Discovery, 2001*, pp. 241–252.

[3] D. Dubois, H. Prade. Gradual inference rules in approximate reasoning. *Proc of the Int. Conf. on Fuzzy Systems, 1992*, vol.61, pp. 103–122.

[4] E. Hüllermeier. Association rules for expressing gradual dependencies. *Principles of Data Mining and Knowledge Discovery , 2002*, vol. 2431, pp. 200–211.

[5] F. Berzal, J. C. Cubero, D. Sanchez, M. A. Vila, J.M. Serrano. An alternative approach to discover gradual dependencies. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2007*, vol. 15, pp. 559–570.

[6] A. Laurent, M. J. Lesot, M. Rifqi. Graank: Exploiting rank correlations for extracting gradual itemsets. *Proc. of the 8th Int. Conf. on Flexible Query Answering Systems, 2009*, pp. 382–393.

[7] L. Di Jorio, A. Laurent, M. Teisseire. Fast extraction of gradual association rules: a heuristic based method. *Proc. of the 5th Int. Conf. on Soft Computing as Transdisciplinary Science and Technology, 2008*, pp. 205–210.

[8] L. Di Jorio, A. Laurent, M. Teisseire. Mining frequent gradual itemsets from large data sets. *Advances in Intelligent Data Analysis VIII, 2009*, pp. 297–308.

[9] B. Bouchon-Meunier, A. Laurent, M. J. Lesot, M. Rifqi. Strengthening fuzzy gradual rules through " all the more " clauses. *Proc of the Int. Conf. on Fuzzy Systems, 2010*, pp. 1–7.

[10] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large data sets, *Proc. 20th Int. Conf. on Very Large data sets*, 1994, pp. 487–499.

[11] R. J. Miller and Y. Yang, Association rules over interval data, *Proc. of the 1997 ACM SIGMOD*, 1997, pp. 452–461.

[12] B. Lent, A. N. Swami and J. Widom, Clustering Association Rules, *Proc. of the 13th Int. Conf. on Data Engineering*, 1997, pp. 220–231.

[13] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, Data mining using two-dimensional optimized association rules: Scheme, algorithm and visualisation, *Proc. of the Int. Conf. ACM SIGMOD*, 1996, pp. 12–23.

[14] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, Mining optimized association rules for numeric attributes, *Proc. of the 15th ACM SIGACT-SIGMOD-SIGART*, 1996, pp. 182–191.

[15] C. Nortet, A. Salleb, T. Turmeaux and C. Vrain, Mining Quantitative Association Rules in a Atherosclerosis Data Set, *Proc. of the 6th PKDD*, 2006, pp.495–506.

[16] J. Mata, J. L. Alvarez and J. C. Riquelme, An evolutionary algorithm to discover numeric association rules, *Proc. of ACM symposium on Applied computing*, 2002, pp. 590–594.

[17] C. Marsala and B. Bouchon-Meunier, Fuzzy Partitioning Using Mathematical Morphology in a Learning Scheme, *Proc. of the 5th Int. Conf. on Fuzz-IEEE*, 1996, pp. 1512–1517.

[18] J. Serra, Introduction to mathematical morphology, *Computer Vision, Graphics, and Image Processing*, 1982, vol. 35, no3, pp 283–305.