Accelerating Effect of Attribute Variations: Accelerated Gradual Itemsets Extraction

Amal Oudni, Marie-Jeanne Lesot, Maria Rifqi

Accelerating Effect of Attribute Variations: Accelerated Gradual Itemsets Extraction

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

[1]Sorbonne Universités, UPMC Univ Paris 06, UMR 7606
LIP6, F-75005, Paris, France.
[2]CNRS, UMR 7606, LIP6, F-75005, Paris, France.
{amal.oudni,marie-jeanne.lesot}@lip6.fr
[3]Université Panthéon-Assas - Paris 02, LEMMA, F-75005, Paris, France.
{maria.rifqi}@u-paris2.fr

Abstract. Gradual itemsets of the form "*the more/less A, the more/less B*" summarize data through the description of their internal tendencies, identified as correlation between attribute values. This paper proposes to enrich such gradual itemsets by taking into account an acceleration effect, leading to a new type of gradual itemset of the form "*the more/less A increases, the more quickly B increases*". It proposes an interpretation as convexity constraint imposed on the relation between A and B and a formalization of these accelerated gradual itemsets, as well as evaluation criteria. It illustrates the relevance of the proposed approach on real data.

Keywords: Gradual Itemset, Acceleration, Enrichment, Convexity.

1 Introduction

Information extraction can take many forms, leading to various types of knowledge which are then made available to experts. This paper focuses on gradual itemsets which can be illustrated by the example "*the closer the wall, the harder the brakes are applied*". Initially introduced in the fuzzy implication formalism [1–3], gradual itemsets 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 is associated with the definition of a support to quantify the validity of gradual itemsets and to methods for the identification of the itemsets that are frequent according to these support definitions.

Furthermore, several types of enrichments have been proposed: in the case of categorical or fuzzy data clauses, clauses linguistically introduced by the expression "all the more" lead to so-called strengthened gradual itemsets [9]. They can be illustrated by an example such as "*the closer the wall, the harder the brakes are applied, all the more the higher the speed*". For numerical data, an enrichment by characterization clauses [10] adds a clause linguistically introduced by the expression "especially if": characterized gradual itemsets can be illustrated

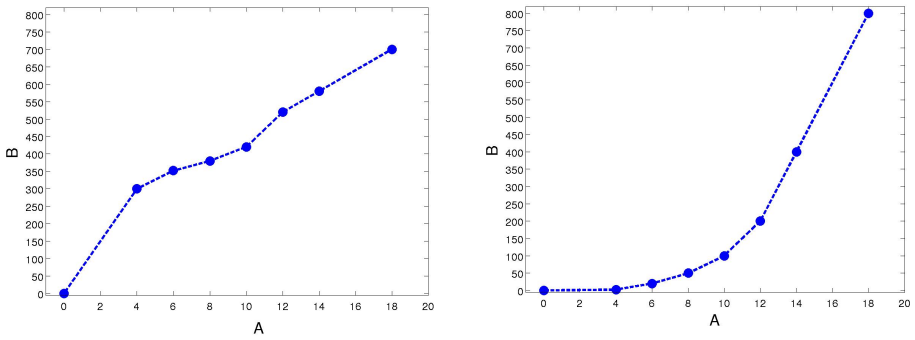


Fig. 1: Two data sets, leading to "*the more A, the more B*" where an acceleration effect is observed for the right data set and not for the left one.

by a sentence as "*the closer the wall, the harder the brakes are applied, especially if the distance to the wall* $\in [0, 50]m$".

In this paper, we consider a new type of enrichment in the case of numerical data, to capture a new type of information: the aim is to express how fast the values of some attributes vary as compared to others, as illustrated by the two data sets represented in Figure 1. In both cases, a covariation constraint is satisfied, which justifies the extraction of the gradual itemset "*the more A, the more B*". However, on the right-hand example, the speed of B augmentation appears to increase, making it possible to enrich the gradual itemset to "*the more A increases, the more quickly B increases*".

This paper addresses the task of extracting such accelerated gradual itemsets. The principle of acceleration is naturally understood as speed variation increase, which can be translated as a convexity constraint on the underlying function associating the considered attributes. This constraint can be modelled as an additional covariation constraint, leading to the definition of a criterion called *accelerated support* to assess the validity of such accelerated gradual itemsets.

The paper is organized as follows: Section 2 recalls the formalism of gradual itemsets and details the existing types of enrichment. Section 3 discusses the proposed interpretation of accelerated gradual itemsets and its formalization. Section 4 defines the criteria proposed for the evaluation of this new type of itemsets. Section 5 illustrates and analyses the experimental results obtained on real data.

2 Typology of Gradual Itemset Enrichments

This section first recalls the notations and definitions of gradual items and itemsets [9, 8] as well as the support definition based on compatible data subsets [8]. It then describes the existing enrichments of gradual itemsets.

2.1 Gradual Itemset Definitions

Let $\mathcal{D}$ denote 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 $M = \{(A_j, *_j), j = 1..k\}$, interpreted as their conjunction. It induces a pre-order, $\preceq_M$, defined as $o \preceq_M o'$ 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.

As briefly recalled in the introduction, there exists several interpretations of gradual itemsets [1–8]. In this paper, we consider the interpretation of 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 $M = \{(A_j, *_j), j = 1..k\}$, $D = \{o_1, ..., o_m\} \subseteq \mathcal{D}$ is a path if and only if there exists a permutation π such that $\forall l \in [1, m-1], o_{\pi_l} \preceq_M o_{\pi_{l+1}}$. Gradual itemsets thus depend on the order induced by the attribute values, not on the values themselves.

Such a path is called *complete* if no object can be added to it without violating the order constraint imposed by M. $\mathcal{L}(M)$ denotes the set of complete paths associated to M. The set of maximal complete paths, i.e. complete paths of maximal length, is denoted $\mathcal{L}^*(M) = \{D \in \mathcal{L}(M) / \forall D' \in \mathcal{L}(M)\ |D| \geq |D'|\}$.

The gradual support of M, $GS_{\mathcal{D}}(M)$, is then defined as the length of its maximal complete paths divided by the total number of objects [7]:

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

2.2 Existing Enrichments

Two enrichment types for gradual itemsets have been proposed, namely through characterization [10] and strengthening [9]. Both are based on a principle of increased validity when the data are restricted to a subset: the gradual support of the considered itemset must increase when it is computed on the data subset only.

More precisely, in the case of characterization [10], the restriction is defined as a set of intervals: characterized gradual itemsets are linguistically of the form "*the more/less A, the more/less B, especially if* $J \in R$", where J is a set of attributes belonging to $A \cup B$ and R is a set of intervals defined for each attribute in J. R defines the data subset, it applies only in the case of numerical data.

In the strengthening case [9], the restriction is defined by a presence, possibly in a fuzzy weighted way, of values required by the strengthening clause: the (fuzzy) data subset only contains objects possessing the required values. Strengthened gradual itemsets are linguistically of the form "*the more/less A, the more/less B, all the more C*", where C is the strengthening clause that consists of values of categorical attributes or fuzzy modalities of fuzzy attributes.

As opposed to the existing enrichments, the peculiarities of the proposed enrichment are mentioned in the following section.

2.3 Characteristics of the Proposed Acceleration Enrichment

The main difference between accelerated gradual itemsets and the previous gradual itemset enrichments comes from the nature of the additional clause: both for characterization and strengthening the enriching clause has a presence semantics, insofar as the additional constraint leads to a data restriction defined by the presence of specific values (in the interval R or in the clause C) on which the itemset validity must increase. On the contrary, as detailed in the next sections, the semantics of the acceleration clause is gradual, depending not on the attribute values but on the order they induce. It thus has the same nature as the considered itemset.

It must be underlined that the accelerated gradual itemsets apply to numerical data, excluding the categorical case.

3 Formalization of Accelerated Gradual Itemsets

This section presents the interpretation and the principle of gradual itemset acceleration, as well as the proposed formalization.

3.1 Principle of Accelerated Gradual Itemsets

As already mentioned in the introduction, Figure 1 represents two data sets with the same cardinality described by two attributes, A (x-axis) and B (y-axis). In both cases, the data sets lead to the same gradual itemset $M = A^{\geq}B^{\geq}$ supported by all data points: the gradual support is 100% in both cases. Now it can be noticed that the covariation between A and B is different: an acceleration effect of B values with respect A values can be observed for the right data set, whereas it does not hold for the left-hand data set.

Accelerated gradual itemsets aim at capturing this difference. It must be underlined that it breaks the symmetry property, distinguishing the cases "the more A, the more quickly B" and "the more B, the more quickly A", whereas the gradual itemset is "the more A, the more B" in both cases.

Mathematically, the acceleration effect corresponds to a convexity property of the function that associates B values to A values, imposing that its graph is "turned up" as illustrated on the right part of Figure 1, meaning that the line segment between any two points on the graph of the function lies above the graph. Convex growth means "increasing at an increasing rate (but not necessarily proportionally to current value)" which is equivalent to desired acceleration effect. Differentiable functions are convex if and only if their derivative is monotonically non-decreasing.

Now, data sets from which accelerated gradual itemsets must be extracted do not give access to the mathematical function relating A and B values, hence its derivative cannot be computed. Therefore we propose to consider a rough discretization, defined as the quotient of the successive differences $\frac{\Delta A}{\Delta B}$ when data are ordered with respect to their A values.

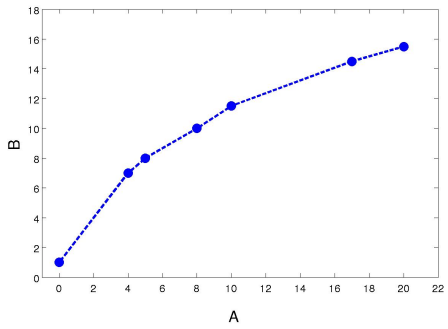


Fig. 2: $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\leq}$ with deceleration effect.

We thus propose to interpret the acceleration effect as an increase of $\left(\frac{\Delta B}{\Delta A}\right)$. It must be noticed that this interpretation does not take into account the shape of the convex function: for instance no difference is made whether the underlying function is quadratic or exponential.

3.2 Formalization

To address the principle presented in the previous section, we propose to formalize an accelerated gradual itemset as a triplet: $A^{*_1}B^{*_2}\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$, where $A^{*_1}B^{*_2}$ represents a gradual itemset, and $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ represents the acceleration clause that compares the variations of B with that of A. $*_1$ determines whether "the more A increases" ($*_1 = \geq$) or "the more A decreases" ($*_1 = \leq$). $*_2$ plays the same role for B. $*_3$ determines whether acceleration or deceleration is considered: $*_3 = \geq$ leads to "the more quickly" and $*_3 = \leq$ leads to "the less quickly" or equivalently "the more slowly" .

This paper focuses on the acceleration effect, i.e. attributes for which values increase "quickly", i.e. the case $*_3 = \geq$. It corresponds to the convex curve case. The case where $*_3 = \leq$ corresponds to a deceleration effect, as illustrated on Figure 2, which can be described as "the more A, the more slowly B increases". It can be noticed that this is equivalent to "the more B increases, the more quickly A increases", i.e. $A^{\geq}B^{\geq}\left(\frac{\Delta A}{\Delta B}\right)^{\geq}$. Thus considering only $*_3 = \geq$ is not a limitation.

3.3 Generalization

The previous definition focuses on the case of itemsets containing two attributes. In the general case, the itemset to enrich may be composed of several attributes, as well as the acceleration clause.

Now the notion of convex function is also mathematically defined for functions depending on several variables, based on properties of their Hessian matrices. Similarly, a discretization based on the available data may be computed

for a given data set, leading to accelerated gradual itemsets made on several attributes, which may be written schematically $M_1 M_2 \frac{\Delta M_2}{\Delta M_1}$.

4 Evaluation Criterion of the Acceleration Effect

An accelerated gradual itemset contains two components, the classical gradual itemset $M = A^{*_1} B^{*_2}$ and the acceleration clause $M_a = \left(\frac{\Delta B}{\Delta A}\right)^{*_3}$. It must therefore be evaluated according to these two components. Its quality is measured both by the classical gradual support as recalled in Equation (1) and an accelerated gradual support that measures the quality of the acceleration, as defined below.

4.1 Order Induced by the Acceleration Clause

The itemset M induces a pre-order on objects as defined in Section 2; the acceleration clause $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ induces a pre-order on pairs of objects denoted $\preceq_{M_a}$: for any o_1, o_2 and o_3

$$(o_1, o_2) \preceq_{M_a} (o_2, o_3) \Leftrightarrow \frac{A(o_2) - A(o_1)}{B(o_2) - B(o_1)} *_3 \frac{A(o_3) - A(o_2)}{B(o_3) - B(o_2)}. \tag{2}$$

where $A(o)$ and $B(o)$ respectively represent the value of attributes A and B for object o.

4.2 Definition of the Accelerated Support

The quality of the candidate accelerated gradual itemset MM_a is high if there exists a subset of data that simultaneously satisfies the order induced by M and that induced by M_a. Therefore the acceleration quality first requires to identify a data subset that satisfies $\preceq_M$. To that aim, the GRITE algorithm [7] can be used to identify candidate gradual itemsets as well as their set of maximal complete support paths $\mathcal{L}^*(M)$.

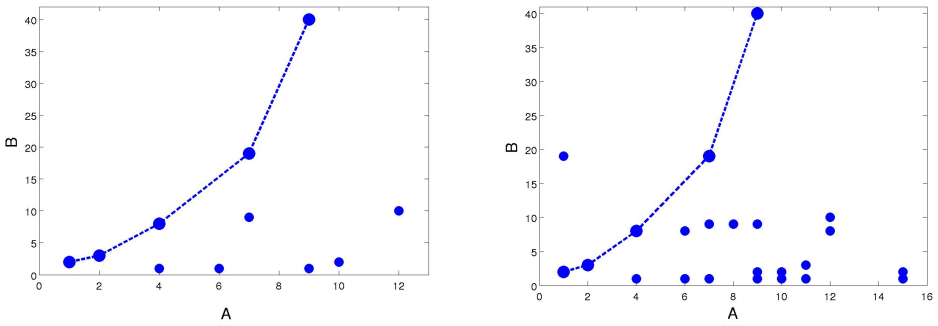
For any $D \in \mathcal{L}^*(M)$, the computation of the accelerated support then consists first in identifying subsets of D so that the constraint $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ is verified simultaneously.

We denote φ the function that identifies a maximal subset of objects from D such that $\forall o_1, o_2, o_3 \in \varphi(D)$, $(o_1 \preceq_M o_2 \preceq_M o_3 \Rightarrow (o_1, o_2) \preceq_{M_a} (o_2, o_3))$

The accelerated gradual support of MM_a is then computed as:

$$GS_a = \frac{1}{|D| - 1} \max_{D \in \mathcal{L}^*(M)} |\varphi(D)| \tag{3}$$

where $|D|$ denotes the size of any maximal complete path in $\mathcal{L}^*(M)$, as, by definition of $\mathcal{L}^*(M)$, they all have the same size. $|D| - 1$ is then the maximal possible value of $\varphi(D)$ and thus the normalizing factor. Indeed, $\left(\frac{\Delta B}{\Delta A}\right)$ does not have the same definition set as classical gradual itemsets: it applies to pairs of successive objects.



(a) $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\geq}$ with high GS. (b) $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\geq}$ with low GS.

Fig. 3: Two data sets for which $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\geq}$ holds with different GS and the same $GS_a = 100\%$.

Combination of the Quality Criterion The classical validity definition is then extended to integrate the condition on GS_a: an accelerated gradual itemset MM_a is valid if $GS \geq s$ and $GS_a(MM_a) \geq s_a$ where s_a is a threshold for the accelerated gradual support and s the threshold of classical gradual support.

It is worth noticing that both GS and GS_a are necessary to assess the quality of an accelerated gradual itemset. Figure 3 illustrates the case of two datasets leading to the same $GS_a = 100\%$ but with different GS: GS equals 45% for the data set on the left and 22% on the right. Indeed GS_a is computed relatively to the path size whereas GS takes into account the total number of points.

When combining the two components, a priority is given to GS: for a given GS level, accelerated gradual itemsets are compared in terms of GS_a.

Computational Cost The computational time of the extraction of accelerated gradual itemsets depends on the number of objects and attributes of the data set, as well as on the gradual support threshold. The experiments described in the next section show that the most expensive step corresponds to the extraction of the basic gradual itemsets and that the step of acceleration clause identification only adds a much smaller computational cost. More precisely, 85% of the total time necessary for the extraction is used in the step of the basic gradual itemset extraction and only 15% is used in the step of the acceleration clause extraction.

5 Experimental Study

This section describes the experiments carried out using the proposed method of accelerated gradual itemset extraction on a real data set. The analysis of the results is based on the number of extracted gradual itemsets and their quality.

5.1 Considered Data

We use 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. They contain 2164 meteorological observations realized during eight days (December 20^{th} to 27^{th} 2013), described by 8 numerical attributes: temperature ($\mathring{C}$), wind chill ($\mathring{C}$), wind run (km), rain (mm), outside humidity (%), pressure (hPa), wind speed (km/hr) and wind gusts measured as high speed (km/hr).

5.2 Results: Extracted Itemsets

Setting a gradual support threshold $s = 20\%$, 153 gradual itemsets are extracted, two of them with 100%. Figure 4 represents the accelerated gradual support of all identified gradual itemsets. It can be observed that itemsets with GS_a below 20% are not numerous and almost 30% have GS_a above 50%. When setting the accelerated support threshold $s_a = 20\%$, represented by the horizontal line on Figure 4, 130 itemsets are considered as enriched by an acceleration clause, which corresponds to more than 85%.

According to the criteria combination with priority discussed in the previous section, the most interesting accelerated gradual itemset is then the one corresponding to point A on the graph. It represents the itemset "the more the wind speed increases, the more quickly the wind run increases: $GS = 100\%$ and $GS_a = 90\%$". This corresponds to an expected result from the proposed definition of accelerated itemsets: the underlying linear relation between these two attributes corresponds to the limit case of acceleration and thus gets a high accelerated support.

The next most interesting itemsets are then the two points in region B in the graph, that respectively correspond to the itemsets

- the more the temperature decreases, the more quickly the rain accumulation increases: $GS = 100\%$ and $GS_a = 32\%$.
- the more the humidity decreases, the more quickly the temperature increases $GS = 94.73\%$, $GS_a = 34\%$.

The middle points in region C in the graph show a trade-off between GS and GS_a. The two ones with highest GS_a correspond to

- the more the wind gusts increase, the more quickly the wind run increases: $GS = 54.9\%$ and $GS_a = 51\%$.
- the more the wind gusts increase, the more quickly the wind speed increases: $GS = 57.3\%$ and $GS_a = 48\%$.

Finally, it can be observed that the majority of extracted gradual itemsets have a gradual support slightly above the threshold 20%, many of them reaching a high accelerated support. Examples with highest GS_a in region D in the graph, include

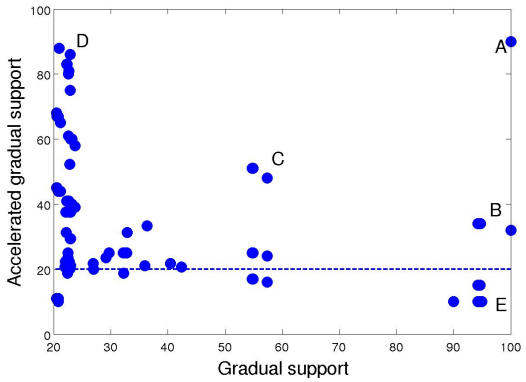


Fig. 4: Gradual support and accelerated gradual support, for each of the 153 extracted gradual itemsets.

- the more the humidity increases, the more quickly the wind run increases: $GS = 22.69\%$ and $GS_a = 81\%$.
- the more the pressure decreases, the more quickly the humidity increases: $GS = 20.93\%$ and $GS_a = 88\%$.
- the more the wind chill increases, the more quickly the temperature increases: $GS = 22.88\%$ and $GS_a = 86\%$.

It is also interesting to look at an example without accelerating effect: the gradual itemset represented by point E corresponds to

- the more the rain accumulation decreases, the more quickly the wind chill increases: $GS = 94.72\%$ and $GS_a = 10\%$.

Accelerated gradual itemsets thus make it possible to extract rich meteorological knowledge from the individual weather station observations.

6 Conclusion and Future Work

In this paper we proposed an approach to enrich gradual itemsets, using an acceleration clause linguistically expressed by the expression "quickly", so as to extract more information summarizing data sets. The extraction of these accelerated gradual itemsets relies on the identification of attributes occurring in the considered gradual itemset for which the speed increase augments compared with other attributes values. The constraint is interpreted in terms of convexity and leads to the definition of a quality criterion to evaluate the acceleration effect.

Ongoing works include complementary experimentations taking into account both computation efficiency (time and memory) and use of other real data where

expert advice can be given on the understanding and interest of extracted accelerated gradual itemsets.

Future works also include the combination of the acceleration effect with other enrichment principles, applied to the acceleration clauses: it would be interesting to identify restriction of the data sets on which the acceleration effect particularly holds. In particular, in the case of meteorological data, restriction induced by a temporal attribute, or by categorical attributes derived from the date, could make it possible to identify accelerated gradual itemsets of the form "the more the temperature increases, the more quickly the rain accumulation decreases, in the summer". Besides, characterization could also allow to remove the ambiguity that may exist when an itemset is extracted with an acceleration and deceleration effect at the same time: a characterization clause would make it possible to identify the subsets of the data where they respectively hold.

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. A. Oudni and M.-J. Lesot and M. Rifqi, Characterisation of gradual itemsets through "especially if" clauses based on mathematical morphology tools, *EUSFLAT*, 2013, pp. 826–833.