GRAANK: Exploiting Rank Correlations for Extracting Gradual Itemsets

Anne Laurent, Marie-Jeanne Lesot, Maria Rifqi

GRAANK: Exploiting Rank Correlations for Extracting Gradual Itemsets

Anne Laurent[1], Marie-Jeanne Lesot[2], and Maria Rifqi[2]

[1] LIRMM-Univ. Montpellier 2 - CNRS UMR 5506,
161 rue Ada, 34 095 Montpellier, France
anne.laurent@lirmm.fr
[2] LIP6 - UPMC - CNRS UMR 7606
104 avenue du Président Kennedy, 75016 Paris, France
{marie-jeanne.lesot,maria.rifqi}@lip6.fr

Abstract. Gradual dependencies of the form *the more A, the more B* offer valuable information that linguistically express relationships between variations of the attributes. Several formalisations and automatic extraction algorithms have been proposed recently. In this paper, we first present an overview of these methods. We then propose an algorithm that combines the principles of several existing approaches and benefits from efficient computational properties to extract frequent gradual itemsets.

Key words: Data Mining, Gradual Dependencies, Gradual Itemsets, Gradual Rules, Ranking Comparison

1 Introduction

Mining digital data sets is one of the key topics addressed in the field of data mining for extracting rules describing the data, their inner trends and exceptions. In this framework, many kinds of patterns and rules can be mined, leading to various pieces of knowledge delivered to experts. In this respect, association rules and sequential patterns are some of the most frequently used patterns that are often provided to the end-users.

In this paper, we focus on gradual dependencies that convey information in the form of attribute covariations, such as *the higher the age, the higher the salary*, or, in a biological application domain, *the higher the expression of gene* G_1 *and the lower the expression of gene* G_2*, the higher the expression of gene* G_3.

Such dependencies resemble gradual rules that have been studied in the context of fuzzy data, and fuzzy implication, in particular for recommendation and command systems [1]: they have the same linguistic form, but their semantics are different. Indeed, fuzzy gradual rules [2] consider the constraints expressed by the rule for each data point individually, requiring that for each point the membership degrees to the modalities involved in the rule satisfy the fuzzy implication, modeled by a residuated implication. Depending on the chosen implication, the gradual rules may include certainty variations, leading to rules such as *the later the waking, the more certain the lateness* [3].

Gradual dependencies take a different approach and consider tendencies across the whole data set, in terms of correlation of the attribute variations, as will be discussed in more details in Section 2. Two kinds of dependencies can be distinguished: a first class considers linguistic variables represented by fuzzy sets and imposes covariation of the membership degrees across all data [4, 5]. They are linguistically expressed as, for example, *the more the age is "middle-aged", the less the number of cars is "low"*, where "middle-aged" and "low" refer to modalities of the linguistic variables age and number of cars respectively. A second, more recent, category directly considers the numerical values of the attributes and applies to attribute covariation on the whole attribute universe [6, 7]. In some works, the focus has also been put on describing the extent to which the degree increases between objects, especially in the framework of temporal digital data [8, 9].

The automatic extraction of gradual dependencies has not received much attention even if it is now gaining interest. As for the association rule extraction, the process consists of two steps: first frequent gradual itemsets, or gradual patterns, are extracted; then, causality relations between the items are looked for. In this paper, we focus on the first step, that aims at identifying frequent gradual itemsets.

Existing techniques make use of different frameworks and formalisations, among which statistic linear regression [4], or level-wise algorithms relying on generalised definitions of support for gradual itemsets [5–7]. Indeed, level-wise approaches, that evaluate $(k+1)$-itemsets knowing the frequent k-itemsets, are very appropriate as the basic property of anti-monotonicity holds when considering gradual patterns: if a gradual pattern containing k attributes (e.g. salary, age) cannot be found in the data set to a sufficient extent, then there is no need to try and find patterns containing these attributes plus other ones. This property allows to design efficient algorithms, as in the case of association rules.

Still the definition of the support of gradual itemsets, i.e. the conditions under which they can be said to occur in the data set, can follow several principles, leading to different definitions and algorithms. In this paper, we first present an overview of the existing approaches, comparing their semantics as well as their properties.

We then show how these approaches can be combined to benefit both from semantic quality and computational efficiency: we follow the definition of gradual dependency based on the notion of order concordance, initially proposed by [5], replacing it in the context of ranking comparison measures. We propose an efficient method, inspired by [7], to compute the corresponding support in a level-wise approach. On one hand the method uses an efficient representation in the form of bitmap matrices, on the other hand, it maintains a concise piece of information that can be queried when navigating from itemsets of one level to the next one. This makes it possible to compute the support without querying the whole data set, which would be both time and memory consuming.

The paper is organized as follows: in Section 2, we recall and compare the main existing approaches allowing users to extract from digital data the hidden

gradual patterns. Section 3 introduces our approach called GRAANK (standing for GRAdual rANKing), including definitions and algorithms, while Section 4 concludes and provides the associated perspectives.

2 State of the art

In this section, after recalling the classic definitions of gradual item, itemset and dependency, as given by [5, 7] for instance, we present the various formalisations and algorithms that have been proposed to automatically extract gradual tendencies from data set. We comment them and underline their differences in terms of interpretation and semantics.

2.1 Gradual items, gradual itemsets and gradual dependencies

Gradual dependencies extraction applies to a data set $\mathcal{D}$ constituted of n objects described by m numerical attributes.

A *gradual item* is defined as a pair made of an attribute and a variation denoted by $\geq$ or $\leq$. If A is an attribute corresponding to the speed of vehicles for instance, $A^{\geq}$ and $A^{\leq}$ are gradual items respectively representing *(speed, more)* and *(speed, less)*. They represent the fact that the attribute values increase (in case of $\geq$) or decrease (in case of $\leq$).

A *gradual itemset* is then defined as a combination of several gradual items, semantically interpreted as their conjunction: for instance $M = A^{\geq}B^{\leq}$ is interpreted as *the more A and the less B*. It imposes a variation constraint on several attributes simultaneously. The number of attributes a gradual itemset involves is called its length.

A *gradual dependency* in turn, denoted $M_1 \rightarrow M_2$, is defined as a pair of gradual itemsets on which a causality relationship is imposed. It can for instance take the form *the faster the speed, then the greater the danger* meaning that a speed increase implies a risk increase: it breaks the symmetry of the gradual itemset in which all items play the same role.

Most existing works about gradual itemsets and gradual dependencies [4, 5] apply to fuzzy data, i.e. data for which the attributes are linguistic variables associated to fuzzy modalities: e.g. a variable representing speed can be associated to 3 modalities, slow, normal, and fast. The data are then described with membership degrees that indicate the extent to which their speeds belong to each modality. A *fuzzy gradual item* is then a triplet made of an attribute, one of its modalities and a variation, such as *(speed, fast, more)*. It is to be understood as "the faster the speed", or more precisely "the higher the membership degree of the speed to fast". It can be represented in the same formalism as the crisp case, introducing one attribute per modality, creating for instance the attributes speedSlow, speedNormal and speedFast whose values are the membership degrees. The fuzzy gradual item can then be written *(speedFast, more)*.

In the following, we use the notation $A^{\geq}$ and $A^{\leq}$ for both crisp and fuzzy data. Moreover, throughout the paper, for any x belonging to the data set $\mathcal{D}$, $A(x)$ denotes the value taken by attribute A for object x.

2.2 Approach based on regression

A first interpretation of gradual dependency expresses it as a co-variation constraint [4]:

Definition 1 (co-variation definition of gradual dependency [4]). *A gradual dependency such as* the more A, the more B *holds if an increase in A comes along with an increase in B.*

In order to identify such relationships, it is proposed in [4] to perform a linear regression analysis between the two attributes. The validity of the gradual tendency is evaluated from the quality of the regression, measured by the normalised mean squared error R^2, together with the slope of the regression line: attribute pairs that are insufficiently correlated are rejected, as well as pairs for which one attribute remains almost constant while the other one increases, which can be detected by a low slope of the regression line.

This definition and this extraction method apply to pairs of attributes. The extension proposed by [4] to longer itemsets considers the case of fuzzy data, for which attributes contain the membership degrees of the data to modalities. It exploits this fuzzy logic framework and the fact that itemsets are interpreted as conjunction of the items they contain: a membership degree to the itemset can be computed using a t-norm, applied to the membership degrees to the items of the considered itemset. The gradual tendency is then understood as a covariance constraint between the aggregated membership degrees. Thus itemsets of length higher than 2 can be handled as itemsets of length 2.

2.3 Formulation as an association rule task

Other works take a different point of view and interpret gradual dependencies as constraints imposed to the order induced by the attributes, and not to their numerical values: in [5] gradual dependencies are considered as generalisations of functional dependencies that replace the equality conditions by variation conditions on the values, leading to the following definition:

Definition 2 (order-based definition of gradual dependency [5]). *A gradual dependency* the more A, the more B *holds if* $\forall x, x' \in \mathcal{D}$, $A(x) < A(x')$ *implies* $B(x) < B(x')$.

It must be underlined that this definition takes into account a causality relationship between the itemsets. It states that the ordering induced by attribute A must be identical to that derived from attribute B. In the case of dependencies such as *the more A, the less B*, the constraint imposes that the orders must be reversed.

In [5] it is proposed to formulate the extraction of such gradual tendencies as the discovery of association rules in a suitable set of transactions obtained from the initial data set $\mathcal{D}$: each pair of objects in the initial data is associated to a transaction in the derived data set $\mathcal{D}'$; items in $\mathcal{D}'$ are defined as A^* ($* \in \{\geq, \leq\}$)

where A are attributes in $\mathcal{D}$. A transaction t in $\mathcal{D}'$ then possesses an item A^* if the pair (x, x') of $\mathcal{D}$ it corresponds to satisfies the constraint imposed by A^*, i.e. $A(x) * A(x')$.

A gradual dependency in $\mathcal{D}$ is then equivalent to a classic association rule extracted from $\mathcal{D}'$. The support of a gradual itemset is thus defined as the proportion of objects couples that verify the constraints expressed by all the gradual items in the itemset [5]:

$$supp(A_1^{*_1}, \dots A_p^{*_p}) = \frac{1}{|\mathcal{D}'|} |\{o = (x, x') \in \mathcal{D}' / \forall j \in [1, p]\ A_j(x) *_j A_j(x')\}| \quad (1)$$

Thus, as the regression approach, this definition also bases the gradual tendencies on correlation between the attributes. However, it considers correlation in terms of the rankings induced by the attributes and not in terms of the values they take. Therefore, it does not rely on any assumption regarding the form of the correlation, e.g. whether it is linear.

Explicitly building the data set $\mathcal{D}'$ to apply a classic frequent itemset extraction algorithm would have too high a computational cost. The authors propose an approximation method, based on the discretization of the attribute values that requires to keep in memory an array of dimension p^k when itemsets of size k are looked for and where p denotes the discretisation level. The computational cost remains high and the experiments are limited to a few attributes.

In an extension [10], it is proposed to take into account the variation amplitude between the object couples: instead of setting binary values in $\mathcal{D}'$ that indicate whether the ordering constraints are satisfied or not, real values are set, depending on the observed values difference: they provide information regarding the extent to which the constraints are satisfied. Fuzzy association rules are then applied to extract information from this data set. This approach bears some similarity to the regression based definition of gradual dependencies that integrates this information through the influence of the regression line slope.

2.4 Approach based on conflict sets

On the basis of Definition 2, a different interpretation is proposed in [6] relying on another definition of support: for an itemset $A_1^{*_1}, \dots, A_p^{*_p}$ the support is defined as the maximal number of rows in $\mathcal{D}$, $\{r_1, \dots, r_l\}$, for which there exists a permutation π such that $\forall j \in [1, l-1], \forall k \in [1, p]$, it holds $A_k(r_{\pi_j}) *_k A_k(r_{\pi_{j+1}})$: denoting $\mathcal{L}$ the set of all such sets of rows, the support is computed as

$$supp(A_1^{*_1}, \dots, A_p^{*_p}) = \frac{1}{|\mathcal{D}|} \max_{L_i \in \mathcal{L}} |L_i| \quad (2)$$

The authors then propose a heuristic to compute this support for gradual itemsets, in a level-wise process that considers itemsets of increasing lengths. It consists in discarding, at each level, the rows whose so-called conflict set is maximal, i.e. the rows that prevent the maximal number of rows to be sorted. It is a heuristic insofar as it performs local choices: choosing a row with smaller

Data set $\mathcal{D}_1$

Object	Att. A	Att. B
o_1	10	12
o_2	12	16
o_3	13	15
o_4	15	17
o_5	20	22

Data set $\mathcal{D}_2$

Object	Att. A	Att. B
o_1	10	12
o_2	12	23
o_3	13	15
o_4	15	17
o_5	20	22

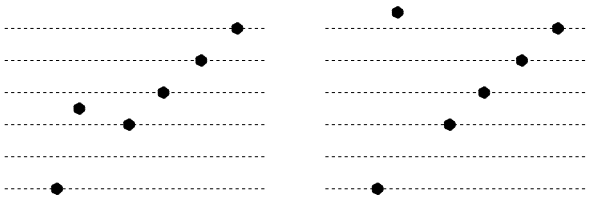

Fig. 1. Two data sets illustrating the influence of deviation amplitude: in both cases, object o_2 contradicts the itemset *the more A, the more B*, also denoted $A^{\geq}B^{\geq}$; for data set $\mathcal{D}_2$, its deviation amplitude is much higher.

conflict set may be suboptimal for a given level but lead to better results at the next level.

It must be underlined that this definition of support gives rise to a major difference with the previous gradual itemsets extraction methods, regarding the data that do not satisfy the itemset: in both the approaches based on regression and on classic association rules, the amplitude of their deviation with the expected behavior is taken into account. This is illustrated on figure 1: in both data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, the object o_2 prevents the itemset *the more A, the more B* to be true, but in the second case, the distortion it leads to is much higher. In other terms, it is more an exception than for $\mathcal{D}_1$.

For the regression based extraction process, this difference is reflected in the regression correlation that is lower for $\mathcal{D}_2$ than for $\mathcal{D}_1$. In the association rule based approach, o_2 lowers the support of the itemset in $\mathcal{D}_2$, because it leads to a high number of data pairs that do not satisfy the gradual itemset, namely $(o_2, o_3), (o_2, o_4)$ and (o_2, o_5). In data set $\mathcal{D}_1$ only the pair (o_2, o_3) contradicts the itemset. Therefore for both methods the itemset $A^{\geq}B^{\geq}$ has a better score for data set $\mathcal{D}_1$ as for $\mathcal{D}_2$.

On the contrary, in the conflict set approach, both $\mathcal{D}_1$ and $\mathcal{D}_2$ lead to the same result: in both cases, it is sufficient to delete point o_2 to obtain a perfect ordering of the data. This shows that gradual itemsets can follow different semantics, the choice between them depends on the considered application.

2.5 Approach based on the precedence graph

In [7], the authors consider the same definition as the one proposed in the conflict set approach (Eq. (2)), and propose a very efficient method based on precedence

graphs, named GRITE for GRadual ITemset Extraction: the data are represented through a graph whose nodes are defined as the objects in the data, and the vertices express the precedence relationships derived from the considered itemset. Moreover, the graph is represented by its adjacency matrix, in a bitmap form: for an itemset $A_1^{*_1}, \dots A_p^{*_p}$, the coefficient corresponding to the object pair (x, x') is 1 if $\forall j \in [1, p]\ \ A_j(x) *_j A_j(x')$, 0 otherwise. The support of the considered itemset can then be obtained as the length of the maximal path in the graph.

The relevance of this approach comes from its very high efficiency to generate gradual itemsets of length $p + 1$ from itemsets of length p: indeed it holds that if s is an itemset generated using s' and s'', its matrix $M_s = M_{s'} \& M_{s''}$ where & is the bitwise AND operation.

3 The proposed GRAANK approach

In this section we propose an algorithm that combines the principles of several existing approaches and benefits from efficient computational properties: we consider the gradual itemset definition used in the association rule formulation [5], see also Section 2.3, and propose an algorithm that exploits the bitmap representation used in the GRITE algorithm [7], see also Section 2.5.

More precisely, we consider the framework of Definition 2 that evaluates gradual tendency in terms of ranking correlation, and the support definition given in Eq. (1). We first interpret it in terms of ranking correlation and then describe the approach proposed to compute the support, and present the derived algorithm.

3.1 Rank correlation measures

Definition 2 directly relates the gradual dependencies extraction task to the framework of ranking comparison: the problem is to compare the rankings induced by all attributes involved in the itemsets, and to compute their degree of agreement, or of correlation.

The problem of rank correlation has been extensively studied by statisticians, and several measures have been proposed, distinguishing between two ranks and multiple rank comparison. Regarding ranking pairs, the most used measures are the Spearman correlation and the Kendall's tau. The latter is of particular interest for gradual itemsets, as its definition matches Eq. (1): given n objects to be ranked, and σ_k, $k = 1, 2$ two rankings where $\sigma_k(x)$ gives the rank of object x in σ_k ranking, the Kendall's tau relies on the definition of concordant and discordant pairs: *concordant pairs* (i, j) are pairs for which the rankings agree, i.e. either $\sigma_1(i) \leq \sigma_1(j)$ and $\sigma_2(i) \leq \sigma_2(j)$, or $\sigma_1(i) \geq \sigma_1(j)$ and $\sigma_2(i) \geq \sigma_2(j)$. Non concordant pairs are called *discordant pairs*. The Kendall's tau is then defined as the proportion of discordant pairs, i.e. the frequency of pair-wise inversions. It is to be noted that the support definition given in Eq. (1) equals the proportion of concordant pairs.

For multiple rank correlation, the mathematical definitions aim at answering significance tests, so as to determine whether the differences between the rankings are significant. For instance Kendall, together with Babington Smith [11] proposed the W coefficient: denoting m the number of rankings to be compared, it is based on the observation that in the case of perfect agreement, the set of values taken by the sums of ranks across the rankings for each object is exactly $m, 2m, \ldots nm$. On the contrary, if the rankings are not correlated at all, the rank sums all take the same value, that equals $n(n-1)/2m$. Therefore, the W coefficient measures the agreement between rankings as the variance of the average rankings for each object. It is normalised by the variance obtained in the ideal case where all rankings are identical. This coefficient is then studied from a statistical point of view, to establish its properties and its distribution, in particular for the definition of significance tests. Moreover equalities between the W coefficient and other approaches to multiple rank correlation, in particular the computation of the average Spearman criterion or the Friedman test, are established, underlining their relationships.

Now such criteria, despite their theoretical properties, cannot be applied to gradual itemset quality evaluation because of computational problems: they do not possess monotony properties that would allow to efficiently prune the set of candidate gradual itemsets when going from one level to the next one. More precisely, for a given set of rankings, the addition of a new ranking in the comparison can lead to increase or decrease the W coefficient. This means that even if an itemset is rejected because of a too low W coefficient, it could be necessary to consider longer itemsets containing it, which would lead to too high a computational complexity.

On the contrary, the ranking comparison induced by the support as defined in Eq. (1), even if it does not possess statistical properties to perform significance tests, is anti-monotonous: the addition of a new item can only decrease the support value. Thus in the following we measure the multiple ranking agreement as the proportion of data pairs that simultaneously satisfy the constraints imposed by all attributes involved in the considered itemsets.

3.2 Support computation

The question then arises how to efficiently compute this quantity. We propose an approach that does not need to perform an approximation as the association rule based method [5]. Following the classic lines, it consists in a level-wise methodology that identifies the relevant gradual itemsets of length $k+1$ from those obtained at level k, the relevance being defined as a support value higher than a user-defined threshold.

To illustrate this section, we consider the data set presented in Table 1 that contains information about 4 persons, regarding their age, salary and the number of their granted loans. Table 2 contains the support values, as defined in Eq. (1), for several itemsets, as well as the list of concordant data couples for each itemset that justifies the computation of the support.

Table 1. Data set example.

Name	A: Age	S: Salary	C: Loans
p_1	22	1200	4
p_2	28	1300	3
p_3	24	2200	5
p_4	35	1850	2

Table 2. Support and list of concordant couples for some gradual itemsets, computed for the data set described in Table 1. To make the table more readable, in the lists of concordant couples, we use the notation (i, j) to denote (p_i, p_j).

Itemset	List of concordant couples	Support
$A^{\geq}S^{\geq}$	$\mathcal{C}_{A^{\geq}S^{\geq}} = \{(1,2)(1,3)(1,4),(2,4)\}$	4/6
$A^{\geq}C^{\leq}$	$\mathcal{C}_{A^{\geq}C^{\leq}} = \{(1,2)(1,4)(2,4)(3,2)(3,4)\}$	5/6
$S^{\geq}C^{\leq}$	$\mathcal{C}_{S^{\geq}C^{\leq}} = \{(1,2)(1,4)(2,4)\}$	3/6
$A^{\geq}S^{\geq}C^{\leq}$	$\mathcal{C}_{A^{\geq}S^{\geq}C^{\leq}} = \{(1,2)(1,4)(2,4)\}$	3/6

List of concordant couples The support value contains aggregated information as it reduces the set of concordant pairs to its cardinality. It can be observed that it is not possible to compute the support for longer itemsets from shorter ones: from the total numbers of pairs that are concordant for two itemsets of length k, the number of pairs concordant for the joint of these sets cannot be derived. Indeed, one cannot determine whether a given object pair is concordant for both itemsets, or for only one of them, and whether it will remain concordant in the joint itemset. Therefore it is necessary to keep, for any itemset, the list of the object pairs that satisfy all the constraints expressed by the involved items, i.e. the list of concordant pairs.

In order to take into account information regarding the sense of variation of the attributes, i.e. to distinguish between $A^{\leq}$ and $A^{\geq}$, we consider object couples instead of pairs, dissociating the two cases of concordance defined in the Kendall's tau: we keep the information whether the couple (i, j) or the couple (j, i) is concordant.

The support then equals the length of the concordant couple list, divided by the total number of object pairs: the latter equals the maximal length of the list, obtained in case of identical rankings. Table 2 contains these lists for the itemsets considered as examples.

List aggregation Keeping such lists of concordance can be compared to the conflict set approach proposed by [6] (see Section 2.4) in which lists of discordant pairs are handled. The difference is that these lists are considered for each data

$A^{\geq}S^{\geq}$:

	1	2	3	4
1	-	1	1	1
2	0	-	0	1
3	0	0	-	0
4	0	0	0	-

$A^{\geq}C^{\leq}$:

	1	2	3	4
1	-	1	0	1
2	0	-	0	1
3	0	1	-	1
4	0	0	0	-

$S^{\geq}C^{\leq}$:

	1	2	3	4
1	-	1	0	1
2	0	-	0	1
3	0	0	-	0
4	0	0	0	-

$A^{\geq}S^{\geq}C^{\leq}$:

	1	2	3	4
1	-	1	0	1
2	0	-	0	1
3	0	0	-	0
4	0	0	0	-

Fig. 2. Binary matrices representing the sets of concordant object pairs for some gradual itemsets, computed for the data set described in Table 1.

point in [6], whereas we propose to attach them to itemsets. The interest of such higher level concordant lists is that they provide an efficient and exact method to generalise gradual itemsets of length k to itemsets of length $k+1$.

Indeed, the list of concordant couples for a gradual itemset s generated using two gradual itemsets s' and s'' is obtained as the intersection of their lists, as illustrated in Table 2 for the considered example.

Formally as s is generated from s' and s'', it only differs from s' by one item that belongs to s'' (and reciprocally):without loss of generality if $s = A_1^{*_1} \dots A_{k+1}^{*_{k+1}}$, denoting $B = A_k^{*_k}$ and $C = A_{k+1}^{*_{k+1}}$ then $s' = A_1^{*_1} \dots A_{k-1}^{*_{k-1}} B$ and $s' = A_1^{*_1} \dots A_{k-1}^{*_{k-1}} C$. Thus an object couple that satisfies all gradual items $A_j^{*_j}$, B and C contained in s' and s'', also satisfies all items in s; reciprocally, if it satisfies all items in s, it satisfies all items contained in s' and s''.

Bitwise representation The problem is then to design an efficient method to store and handle the lists of concordant couples. To that aim, we propose to exploit a bitwise representation as used by [7]: it consists in defining a matrix for each considered itemset, that indicates for each couple of data, whether it is concordant or not: for an itemset $A_1^{*_1}, \dots A_p^{*_k}$, the value corresponding to the object couple (x, y) is 1 if $\forall j \in [1, k]\, A_j(x) *_j A_j(y)$, 0 otherwise. These matrices are illustrated on Figure 2 for the considered example.

On one hand, this representation makes it very easy to go from itemsets of length k to itemsets of length $k+1$: the intersection of the concordance lists equals the bitwise AND operation between the corresponding matrices. On the other hand, the support of the itemset is simply obtained as the sum of the elements of the matrix, divided by the total number of pairs of objects.

3.3 The GRAANK algorithm

The proposed algorithm thus follows the principle of the APRIORI algorithm, modifying the step of candidate itemset evaluation that is performed using the

efficient support computation described in the previous section. More precisely, the algorithm works as follows:

1. Initialization ($k = 1$): for all attributes A, build the concordance matrices for $A^{\geq}$ and $A^{\leq}$.
2. Candidate gradual itemset generation at level $k + 1$:
   apply the APRIORIgen algorithm to generate candidates from the k-itemsets, computing their concordance matrix as the logical AND of the concordance matrices of the joined k-itemsets.
3. Candidate evaluation:
   (a) for all candidate itemsets, compute their support, as the sum of their concordance matrices divided by $n(n-1)/2$ where n is the number of objects.
   (b) discard candidates whose support is lower than the user-defined threshold.
4. Iterate on step 2 and 3 until the generation step does not provide any new candidate.

It is to be underlined that this algorithm is very efficient in several aspects: the support computation does not require any counting operation performed on the data set, and can be deduced from information of the previous level. Moreover this information can be handled in an efficient manner too, thanks to the bitwise representation of the concordance matrices.

These advantages are similar to that of the algorithm proposed by [7]. The difference with the latter comes from the candidate evaluation step: the search of the longest path in the matrix is replaced by the simple sum of its components. On one hand, this lowers the computational complexity of the approach. On the other hand, this difference, that appears to be a minor one, actually deeply modifies the semantics of the induced gradual itemsets: it makes it possible to take into account the amplitude of the distortion for data that do not satisfy the gradual itemsets. It offers an alternative interpretation of gradual constraints, whose validity depends on the application and should be evaluated by the user. For the example described in Table 1, the definition of support as longest path leads to a value of 3/4 for the gradual itemset $A^{\geq}S^{\geq}C^{\leq}$: it is sufficient to suppress one object, p_3, so that the ordering constraint holds. Using the support definition as length of the concordant couples list, the deviation amplitude is taken into account for this point. Now in this case, p_3 is very young for his high salary, while he has a high number of loans as compared to the other persons, which makes him a very exceptional case. This significantly decreases the gradual itemset support to 1/2. This illustrates the differences between the support definitions.

The proposed approach thus offers an efficient implementation of the support definition proposed by [5], interpreting it in the framework of rank comparison. From a computational point the view, the algorithm benefits from the efficiency of the approach based on binary matrices [7]. Moreover, as the cost of the matrix sum is lower than that of searching the longest path in the precedence graph, even if an efficient algorithm is proposed by [7], its complexity is even lower.

4 Conclusion

In this paper, we propose an original approach called GRAANK for extracting gradual patterns. This approach integrates complementary paradigms: the definition of gradual itemset support based on rank correlation, the efficiency of level-wise approaches and that of bitmap representation. We provide the necessary formal definitions, together with the associated algorithms.

Beside extensive experimentations including both computation efficiency (time and memory) and semantics (relevance and comparison of the extracted patterns), further works include the study of other optimizations in order to improve the efficiency of our approach. Moreover, we aim at studying how causality links (rules) can be extracted, and how temporality can be handled, for instance to manage databases describing digital records taken by sensors at several time periods. Finally, we aim at studying how our approach, essentially designed to point out the main tendencies from a digital database, can also be considered for pointing out outliers.

References

1. Galichet, S., Dubois, D., Prade, H.: Imprecise specification of ill-known functions using gradual rules. Int. Journal of Approximate Reasoning **35** (2004) 205–222
2. Bouchon-Meunier, B., Dubois, D., Godó, L., Prade, H.: Fuzzy sets and possibility theory in approximate and plausible reasoning. In Bezdek, J., Dubois, D., Prade, H., eds.: Fuzzy sets in approximate reasoning and information systems. Kluwer Academic Publishers (1999) 15–190
3. Dubois, D., Prade, H.: Gradual inference rules in approximate reasoning. Information Sciences **61**(1-2) (1992) 103–122
4. Hüllermeier, E.: Association rules for expressing gradual dependencies. In: Proc. of the 6th European Conf. on Principles of Data Mining and Knowledge Processing, PKDD'02, Springer-Verlag (2002) 200–211
5. Berzal, F., Cubero, J.C., Sanchez, D., Vila, M.A., Serrano, J.M.: An alternative approach to discover gradual dependencies. Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS) **15**(5) (2007) 559–570
6. Di Jorio, L., Laurent, A., Teisseire, M.: Fast extraction of gradual association rules: a heuristic based method. In: Proc. of the IEEE/ACM Int. Conf. on Soft Computing as a Transdisciplinary Science and Technology, CSTST'08. (2008)
7. Di Jorio, L., Laurent, A., Teisseire, M.: Mining frequent gradual itemsets from large databases. In: Proc. of the Int. Conf. on Intelligent Data Analysis, IDA'09. (2009)
8. Fiot, C., Masseglia, F., Laurent, A., Teisseire, M.: Ted and Eva: Expressing temporal tendencies among quantitative variables using fuzzy sequential patterns. In: Fuzz'IEEE. (2008)
9. Fiot, C., Masseglia, F., Laurent, A., Teisseire, M.: Evolution patterns and gradual trends. Int. Journal of Intelligent Systems (2009)
10. Molina, C., Serrano, J.M., Sánchez, D., Vila, M.: Measuring variation strength in gradual dependencies. In: Proc. of the European Conf. EUSFLAT'07. (2007) 337–344
11. Kendall, M., Babington Smith, B.: The problem of m rankings. The annals of mathematical statistics **10**(3) (1939) 275–287