Mining Frequent Seasonal Gradual Patterns

Jerry Lonlac, Arnaud Doniec, Marin Lujak, Stéphane Lecoeuche

Mining Frequent Seasonal Gradual Patterns

Jerry Lonlac, Arnaud Doniec, Marin Lujak, Stephane Lecoeuche
Research Center, IMT Lille Douai, University of Lille, 59500 Douai, France
{jerry.lonlac,arnaud.doniec,marin.lujak,stephane.lecoeuche}@imt-lille-douai.fr

***Abstract*—Mining frequent episodes aims at recovering sequential patterns from temporal data sequences, which can then be used to predict the occurrence of related events in advance. On the other hand, gradual patterns that capture co-variation of complex attributes in the form "When X increases/decreases, Y increases/decreases" play an important role in many real world applications where huge volumes of complex numerical data must be handled. More recently, they have received attention from the data mining community for exploring temporal data and methods have been defined to automatically extract gradual patterns from temporal data. However, to the best of our knowledge, no method has been proposed to extract gradual patterns that always appear at the identical time intervals in the sequences of temporal data, despite the knowledge that such patterns may bring in certain applications. In this paper, we propose to extract co-variations of periodically repeating attributes from the sequences of temporal data that we call *seasonal gradual patterns*. We discuss the specific features of these patterns and propose an approach for their extraction by exploiting motif mining algorithms in a sequence, and justify its applicability to the gradual case. Illustrative results obtained from a real world data set are described and show the interest for such patterns.**

***Index Terms*—data mining, gradual patterns, temporality, seasonal tendencies**

I. Introduction

Due to the abundance of data collection devices and sensors, numerical data are ubiquitous and produced in increasing quantities. They are produced in many domains including e-commerce, biology, medicine, environment and ecology, telecommunications, and system supervision. In recent years, the analysis of numerical data has received attention from the data mining community and methods have been defined for dealing with such data. These methods have allowed to automatically extract different kinds of knowledge from numerical data express under form of patterns such as quantitative item set/association rules [1]–[3], interval patterns [4] and gradual pattern mining [5]. More recently, gradual patterns that model frequent co-variations between numerical attributes, such as "the more experience, the higher the salary, and the lower the free time" aroused great interest in a multitude of areas. For example, in medicine, gradual patterns make it possible to capture the correlations between memory and feeling points from the *Diagnostic and statistical manual of mental disorders* [6]; in biology, they help to analyze genome data for discovering correlations between genomic expressions; in health, where researchers often seek to explain the differences between micro sequences (DNA, gene or protein) according to clinical characteristics (grade of a tumor, number of recurrences, etc.); in financial markets, where one would like to discover co-evolution between financial indicators, or in marketing for analyzing client databases. Several works have addressed the mining of gradual patterns and different algorithms have been designed for discovering gradual patterns from numerical data (e.g., [5], [7]–[11]).

Most of these algorithms use data mining techniques to extract gradual patterns. However, they are not relevant for extracting gradual patterns in certain application domains where numerical data present particular forms (e.g., temporal, stream, relational, or noisy data). So, some recent works have instead focused on extracting variants of gradual patterns on the numerical data supplied with specific constraints for expressing another kind of knowledge. For instance, in [12], an approach based on B-Trees and OWA (Ordered Weighted Aggregation) operator [13], [14] is proposed to mine data streams for gradual patterns. [15] propose the relational gradual pattern concept, which enables to examine the correlations between attributes from a graduality point of view in multi-relational data.

Fuzzy gradual patterns are revisited in [16] for noisy data where it is often hardly possible to compare attribute values, either because the values are taken from noisy data, or because it is difficult to consider that a small difference between two values is meaningful. An example of a fuzzy gradual pattern could be expressed as "the closer the age of an employee to 46, the higher his/her income". Recently, in [17], the authors proposed an approach to extract gradual patterns in temporal data with an application on paleoecological databases to grasp functional groupings of coevolution of paleoecological indicators that model the evolution of the biodiversity over time. [18] introduce a generic method for extracting and analyzing gradual patterns in spatial data at several levels of granularity. The authors apply their method on the Health data to measure potentially avoidable hospitalization related with both societal and financial issues in public policies. More recently, in [19] propose fuzzy temporal gradual patterns to integrate the fact that a temporal lag may exist between changes in some attributes and their impact on others. These fuzzy temporal gradual patterns alow to detect the cases of relevant correlations between the attributes of a database whose changes in the value one attribute causes a ripple effect on other attributes with respect to time.

The works described above shed light to the interest for gradual patterns in many domains and show that according to the requirements linked to data structures, new methods are needed to extract the forms of gradual patterns expressing typical knowledge to the data context. Let us mention that the quantity of gradual patterns generated from a database can be

high, which makes it hard for exploitation by a user. Moreover, each gradual pattern extracted is provided with a subsequence set of objects, called extension, supporting it. These extensions associated to gradual patterns, whose number can be important, are often useful in several application domains for deriving new and relevant knowledge to the expert. To this end, recently [20] propose a Sequential Pattern Mining based approach for efficient extraction of frequent gradual patterns with their corresponding sequence of tuples. For instance, in [18], the authors show on geographical data that the analysis of the different sequence of objects associated to the gradual patterns allows to identify how an object participates in the associations between attribute variations. Therefore, it is often important for an algorithm to return a reasonable quantity of patterns with their associated extension. This reinforced the need to design the pattern mining algorithms adapted to the issues raised by the considered data. In this framework, recall that the approach proposed by [17] only extracts the gradual patterns associated to the sequence of consecutive objects and that follow a temporal order. Although this approach deals with temporally annotated numerical data, it can be only used on the data constituted as a single valued sequence. When considering the databases constituted as many ordered value sequences, state of the art gradual pattern mining algorithms do not allow extracting another kind of knowledge that is typical of sequential data.

Starting from all these observations, in this paper we propose an approach for extracting seasonal gradual patterns, (the latter term proposed by [17]) but always associated to the same sub-sequence of objects in all value sequences. We justify the interest for such patterns in the e-commerce domain where it is very important to understand seasonal patterns. In fact, the knowledge brought by such patterns can be used in logistics for decision marking in, e.g., inventory and supply chain management. Our approach can be seen as an extending gradual patterns in the temporal context via the seasonality notion such that we include the seasonal correlations between attributes. Let us recall that gradual patterns can be compared to fuzzy gradual rules that have first been used for command systems some years ago [21]. In fact, fuzzy gradual rules that refers to patterns like "the closer the wall, the stronger the brake force" are expressed in the same way as the gradual patterns, but they were not discovered automatically from data.

II. Preliminaries

This section formally describes the problem of mining frequent gradual itemsets (patterns) in a numerical database. We present some of the state-of-the-art approaches proposed to automatically extract such patterns. Finally, the problem of enumerating all motifs in a sequence (EMS) that we exploit for our approach is also presented.

A. Gradual patterns mining problem

The problem of mining gradual patterns consists in mining attribute co-variations in a numerical dataset of the form *"The more/less X, . . . , the more/less Y"*. We assume herein that we are given a database Δ containing a set of objects $\mathcal{T}$ that defines a relation on an attribute set $\mathcal{I}$ with numerical values. Let $t[i]$ denote the value of attribute i over object t for all $t \in \mathcal{T}$.

TABLE I
Database Δ_1

tid	age	salary	cars	loans
t1	22	1000	2	4
t2	24	1200	3	3
t3	28	1850	2	5
t4	20	1250	4	2
t5	18	1100	4	2
t6	35	2200	4	2
t7	38	3200	1	1
t8	44	3400	3	6
t9	52	3800	3	3
t10	41	5000	2	7

In Table I, we give an illustrative example of a numerical database built over the set of attributes $\mathcal{I} = \{age, salary, cars, loans\}$.

Each attribute will hereafter be considered twice: once to indicate its increasing and once to indicate its decreasing, using the $\uparrow$ and $\downarrow$ variation symbols, where $\uparrow$ stands for increasing and $\downarrow$ stands for decreasing variation. In the following, for all $t, t' \in \mathcal{T}$ and for all $i \in \mathcal{I}$, we denote "$t[i] \uparrow t'[i]$" (respectively "$t[i] \downarrow t'[i]$") to mean that the value of attribute i increases (respectively decreases) from t to t'.

Definition 1 (Gradual item): Let Δ be a data set defined on a numerical attribute set $\mathcal{I}$. A gradual item is defined under the form i^*, where i is an attribute of $\mathcal{I}$ and $* \in \{\uparrow, \downarrow\}$.

If we consider the numerical database of Table I, $age^{\uparrow}$ (respectively $age^{\downarrow}$) is a gradual item meaning that the values of attribute *age* are increasing (respectively decreasing).

A gradual pattern is thus defined as follows:

Definition 2 (Gradual pattern): A gradual pattern $g = (i_1^{*1}, ..., i_k^{*k})$ is a non-empty set of gradual items. A k-itemset is an itemset containing k ($k > 1$) gradual items.

For example, $g_1 = \{age^{\uparrow}, salary^{\uparrow}\}$ is a gradual pattern (2-itemsets) meaning that *"the higher the age, the higher the salary"*.

A gradual itemset imposes a variation constraint on several attributes simultaneously. The length of a gradual itemset is equal to the number of gradual items that it contains.

The support (frequency) of a gradual pattern amounts to the extent to which a gradual pattern is present in a given database. Several support definitions have been proposed in the literature (e.g., [5], [7], [8], [22]), showing that gradual patterns can follow different semantics. In [7] the computation of the support of gradual patterns is based on linear regression. [22] and [8] consider the proportion of couples of tuples that verifies the constraints expressed by all the gradual items of the pattern while in [5], the support is defined as the size of the longest sequence of tuples supporting the gradual pattern. More recently, [17] define the support of a gradual pattern respecting the temporal order as the proportion of couples of consecutive tuples supporting the gradual pattern. In this paper,

we adopt this last definition of support, which is more adapted to our issue. To define this support, we introduce the following definitions:

Definition 3 (Gradual tuple motif): Let $g = (i_1^{*_1}, \ldots, i_k^{*_k})$ be a gradual itemset and $M = t_1 t_2 \ldots t_n$ be a motif of consecutive tuples. M is gradual with respect to g if for all p such that $1 \leq p \leq k$ and for all j such that $1 \leq j < n$, the following constraint is satisfied:

$$t_j[i_p] \ *_p \ t_{j+1}[i_p] \tag{1}$$

Considering the database of Table I, $M_1 = t_1 t_2 t_3$ is a gradual tuple motif with respect to $g_1 = \{age^{\uparrow}, salary^{\uparrow}\}$.

It is important to note that there may be several gradual tuple motifs respecting g.

Definition 4 (Maximal gradual tuple motif): Let $g = (i_1^{*_1}, ..., i_k^{*_k})$ be a gradual itemset and M a gradual tuple motif respecting g. M is maximal if for any motif M' respecting g, $M \not\subset M'$.

For instance, when considering the data from Table I and the gradual pattern g_1. The gradual tuple motif $M_2 = t_6 t_7 t_8 t_9$ is not maximal with respect to g_1 because the motif $M_3 = t_5 t_6 t_7 t_8 t_9$ is gradual with respect to g_1 and contains M_2. M_3 is a maximal gradual tuple motif with respect to g_1.

Definition 5 (Cover): Let g be a gradual itemset of a numerical database Δ. We define $Cover(g, \Delta)$ as the set of maximal gradual tuple motifs in respect to g in Δ.

Considering the database of Table I and the previous gradual itemset g_1, $Cover(g_1, \Delta_1) = \{t_1 t_2 t_3, t_5 t_6 t_7 t_8 t_9\}$.

Definition 6 (Gradual tuple motif sequence): Let g be a gradual itemset of a database Δ and $\mathring{f}$ be a function such that $\mathring{f}(M_1, \ldots, M_n) = M_1 \circ \ldots \circ M_n$ where $M_j (1 \leq j \leq n)$ is a gradual tuple motif. Then we define the gradual tuple motif sequence of g in Δ noted M_g^{Δ} as $M_g^{\Delta} = \mathring{f}(cover(g, \Delta))$.

A gradual tuple motif sequence is just a concatenation of gradual tuple motifs. Referring back to our previous example from Table I, we have $M_{g_1}^{\Delta_1} = t_1 t_2 t_3 \circ t_5 t_6 t_7 t_8 t_9$.

III. Problem statement

In this section, we describe the type of data used, the notion of seasonal gradual patterns, and its examples.

A. Temporal data sequences

Our approach finds its application on a numerical database Δ constituted of temporal data sequences. More precisely, the database Δ consists of object sequences $\mathcal{S} = \langle S^1, \ldots, S^n \rangle$ described by the set of numerical attributes $\mathcal{I} = \{i_1, \ldots, i_k\}$, with $S^j = \{d_1, \ldots, d_l\}$, a set of periods considered. Table II is an example of temporal data sequences which gives information about customer purchases for a e-commerce website on three purchase cycles (sequences) S^1, S^2, S^3. Each sequence contains the data for eight dates $(d_1, \ldots, d_8)$. Without loss of generality, we assume that there are no other purchase dates between two consecutive dates and that the purchases are made continuously between two consecutive cycles.

B. Seasonal gradual patterns

In the case of a single object sequence, a gradual pattern corresponds to the one extracted by [17]. However, in seasonal gradual pattern context, we seek for the gradual patterns respected by the same gradual tuple motifs. To address this issue, we propose the definition of seasonal gradual patterns in which the notion of seasonality is introduced. Let us consider the temporal data sequence of Table II. These data are extracted from a dataset regarding customer orders made at multiples marketplaces.

The goal is to extract frequent co-variations between attribute values that occur frequently in identical periods, e.g. seasonal gradual patterns. We want to extract these patterns with the gradual tuple motifs associated that will represent the seasonality of each pattern. To illustrate our approach, we start by introducing some definitions.

Definition 7: Let Δ be a temporal data sequence over a set of numerical attributes $\mathcal{I} = \{i_1, \ldots, i_k\}$, and of tuple sequences $\mathcal{S} = \langle S^1, \ldots, S^n \rangle$. Given gradual item i^* with $i \in \mathcal{I}$, we define M_{i^*} as $M_{i^*} = \mathring{f}(M_{i^*}^{S^1} \ldots M_{i^*}^{S^n})$.

M_{i^*} is the sequence formed of gradual tuple motifs that respect gradual item i^*.

Example 1: Referring back to the example from Table II, we have: $M_{age^{\uparrow}}^{S^1} = d_1 d_2 d_3 \circ d_5 d_6 d_7 d_8$, $M_{age^{\uparrow}}^{S^2} = d_1 d_2 d_3 d_4 \circ d_7 d_8$, $M_{age^{\uparrow}}^{S^3} = d_1 d_2 d_3 \circ d_4 d_5 d_6 d_7$. Then $M_{age^{\uparrow}} = d_1 d_2 d_3 \circ d_5 d_6 d_7 d_8 \circ d_1 d_2 d_3 d_4 \circ d_7 d_8 \circ d_1 d_2 d_3 \circ d_4 d_5 d_6 d_7$.

Note that a given gradual item i^* corresponds to a unique gradual tuple motif sequence M_{i^*}. Let us now give some basic definitions and notations necessary to introduce our approach.

Let M_{i^*} be a motif sequence of gradual tuples. We denote by $O = \{1 \ldots |M_{i^*}|\}$ the set of positions of the tuples in M_{i^*}.

Definition 8 (Inclusion): A gradual tuple motif $M = t_1 \ldots t_m$ appears in a gradual tuple motif sequence $M_g = s_1 \ldots s_n$ at the position $l \in O$ denoted $M \subseteq_l M_g$, if $\forall j \in \{1 \ldots m\}, t_j = s_{l+j-1}$ and $t_j \neq o$. We note by $\mathcal{L}_{M_g}(M) = \{l \in O | M \subseteq_l M_g\}$ the support of M in M_g. We say that $M \subseteq M_g$ iff $\exists l \in O$ such that $M \subseteq_l M_g$.

Definition 9 (Frequent gradual tuple motif): Let M_g be a gradual tuple motif sequence and M gradual tuple motif. Given a positive number $\theta \geq 1$, called quorum, we say that M is frequent in M_g when $|\mathcal{L}_{M_g}(M)| \geq \theta$.

In the rest of the paper, given a gradual tuple motif sequence M_g, the set of all frequent maximal gradual tuple motifs of M_g for the quorum θ is denoted by $\mathcal{E}_{M_g}^{\theta}$. We now consider a new kind of items that we call seasonal gradual items defined as follows:

Definition 10 (Seasonal gradual item): Let Δ be a temporal data sequence over a set of numerical attributes $\mathcal{I} = \{i_1, \ldots, i_k\}$ and θ a support threshold. A seasonal gradual item with respect to θ is defined under form $i^{(*,m)}$, where $* \in \{\uparrow, \downarrow\}$ and $m \in \mathcal{E}_{M_{i^*}}^{\theta}$.

Example 2: If we consider the temporal data sequence of Table II, $age^{(\uparrow, d_1 d_2 d_3)}$ is a seasonal gradual item with respect to $\theta = 3$ expressing that the values of the attribute age are increasing more frequently on the period ¡¡$d_1 d_2 d_3$¿¿.

TABLE II
CUSTOMER PURCHASES DATABASE: Δ_2

Sid	purchase_timestamp	age (a)	freight_value (f)	payment_installments (pi)	payment_value (pv)
S^1	d1	22	8.72	2	18.12
	d2	24	22.76	3	141.46
	d3	28	19.22	4	179.12
	d4	20	17.20	1	72.20
	d5	18	8.72	1	28.62
	d6	35	27.36	3	175.26
	d7	38	16.05	4	65.95
	d8	44	15.17	4	75.16
S^2	d1	32	16.05	3	35.95
	d2	34	19.77	4	161.42
	d3	36	30.53	5	159.06
	d4	40	16.13	5	114.13
	d5	25	14.23	2	50.13
	d6	23	12.805	2	32.70
	d7	20	13.11	1	54.36
	d8	41	14.05	4	46.45
S^3	d1	28	77.45	3	1376.45
	d2	33	15.10	4	43.09
	d3	38	11.85	6	29.75
	d4	35	16.97	5	62.15
	d5	38	8.96	4	118.86
	d6	44	8.71	5	88.90
	d7	52	7.78	6	17.28
	d8	41	57.58	4	187.57

$age^{(\uparrow, d_5d_6d_7)}$ is not a seasonal gradual item with respect to $\theta = 3$ as the support of $d_5d_6d_7$ in $M_{age^{\uparrow}}$ is equal to 2 (observed on S^1 and S^3, not on S^2).

Definition 11 (Seasonal gradual itemset): A seasonal gradual itemset (pattern) $g = \{i_1^{(*_1,m)}, \ldots, i_k^{(*_k,m)}\}$ is a non-empty set of seasonal gradual items, with $m \in \mathcal{E}^{\theta}_{M_g}$ and θ a support threshold.

Example 3: Consider the temporal data sequence of Table II, $g_1 = \{age^{(\uparrow, d_1d_2d_3)}, payment\_installments^{(\uparrow, d_1d_2d_3)}\}$ is a seasonal gradual pattern meaning that "an increase of age comes along with an increase of $payment\_installments$ more frequently on the period "$d_1d_2d_3$".

Definition 12 (Frequent seasonal gradual patterns mining problem): Let Δ be a temporal data sequence and θ a minimum support threshold. The problem of mining seasonal gradual patterns is to find the set of all frequent seasonal gradual patterns of Δ with respect to θ.

Let us indicate that, in the classical patterns mining framework, the problem of enumerating all motifs possibly interspersed with a wildcard symbol, in a sequence of items that has been extensively investigated in e.g., ([23]–[26]) is related to the frequent seasonal gradual patterns mining problem. In fact, for a gradual item i^*, mining all frequent maximal gradual tuple motifs of M_{i^*} corresponds to the problem of enumerating all maximal frequent motifs with a wildcards symbol in a sequence of tuples.

IV. EXTRACTING SEASONAL GRADUAL PATTERNS

In this section, we describe how to extract seasonal gradual patterns from a numerical temporal data sequence. We first transform the frequent seasonal gradual patterns mining problem into the problem of enumerating all motifs with a wildcard symbol in a tuple sequences database. This transformation is given by the following definition.

Definition 13: Let Δ be a temporal data sequence over a set of numerical attributes $\mathcal{I} = \{i_1, \ldots, i_k\}$. We define $\Gamma(\Delta)$ the tuple sequences database associated to Δ as $\Gamma(\Delta) = \{(i_1^{\uparrow}, M_{i_1^{\uparrow}}), (i_1^{\downarrow}, M_{i_1^{\downarrow}}), \ldots, (i_k^{\uparrow}, M_{i_k^{\uparrow}}), (i_k^{\downarrow}, M_{i_k^{\downarrow}})\}$.

TABLE III
TUPLE SEQUENCES DATABASE $\Gamma(\Delta_2)$ OBTAINED FROM DATABASE Δ_2

Gradual Items	Tuple Sequences
$a^{\uparrow}$	$d_1d_2d_3 \circ d_5d_6d_7d_8 \circ d_1d_2d_3d_4 \circ d_7d_8 \circ d_1d_2d_3 \circ d_4d_5d_6d_7$
$a^{\downarrow}$	$d_3d_4d_5 \circ d_8d_1 \circ d_4d_5d_6d_7 \circ d_8d_1 \circ d_3d_4 \circ d_7d_8$
$f^{\uparrow}$	$d_1d_2 \circ d_5d_6 \circ d_8d_1d_2d_3 \circ d_6d_7d_8d_1 \circ d_3d_4 \circ d_7d_8$
$f^{\downarrow}$	$d_2d_3d_4d_5 \circ d_6d_7d_8 \circ d_3d_4 \circ d_5d_6 \circ d_1d_2d_3 \circ d_4d_5d_6d_7$
$pi^{\uparrow}$	$d_1d_2d_3 \circ d_4d_5d_6d_7d_8 \circ d_1d_2d_3d_4 \circ d_5d_6 \circ d_7d_8 \circ d_1d_2d_3 \circ d_5d_6d_7$
$pi^{\downarrow}$	$d_3d_4d_5 \circ d_7d_8d_1 \circ d_3d_4d_5d_6d_7 \circ d_8d_1 \circ d_3d_4d_5 \circ d_7d_8$
$pv^{\uparrow}$	$d_1d_2d_3 \circ d_5d_6 \circ d_7d_8 \circ d_1d_2 \circ d_6d_7 \circ d_8d_1 \circ d_3d_4d_5 \circ d_7d_8$
$pv^{\downarrow}$	$d_3d_4d_5 \circ d_6d_7 \circ d_8d_1 \circ d_2d_3d_4d_5d_6 \circ d_7d_8 \circ d_1d_2d_3 \circ d_5d_6d_7$

Example 4: For instance, the tuple sequences database associated to database Δ_2 of Table II is given by Table III.

Proposition 1 illustrates the mapping between the set of seasonal gradual items of Δ_2 and the maximal motifs of $\Gamma(\Delta_2)$.

Proposition 1: Let Δ be a temporal data sequence and θ a minimal support threshold (quorum). $g = \{i_1^{(*_1,m)}, \ldots, i_k^{(*_k,m)}\}$ is a seasonal gradual pattern of Δ iff $\forall 1 \leq p \leq k$, $|\mathcal{L}_{M_{i_p^*}}(m)| \geq \theta$ with $Cover(m, \Gamma(\Delta)) = g$. Moreover $Cover(g, \Delta)$ is the set of maximal tuple motifs m of $\Gamma(\Delta)$ with $Cover(m, \Gamma(\Delta)) = g$.

As mentioned above, there are different support definitions for the classical gradual patterns. The support to measure the graduality of seasonal gradual pattern is defined as follows:

Definition 14 (Seasonal gradual pattern support computation): Let Δ be a temporal data sequence and $g = \{i_1^{(*_1,m)}, \ldots, i_k^{(*_k,m)}\}$ be a seasonal gradual pattern of Δ. The support of g can be defined as follows: $Supp(g,\Delta) = \frac{min\{|\mathcal{L}_{M_{i_p^{*_p}}}(m)|, 1 \leq p \leq k\} \times |m|}{|\Delta|}$.

Given a predefined minimal support threshold θ and a seasonal gradual pattern g, we say that g is frequent if its support is greater than or equal to θ. The support definition of seasonal gradual pattern satisfies the classical anti-monotony property.

Proposition 2: Let M be a maximal gradual tuple motif of $\Gamma(\Delta)$ then $g = cover(\mathrm{M}, \Gamma(\Delta))$ is a seasonal gradual pattern in Δ. with $Supp(g,\Delta) \geq \frac{min\{|\mathcal{L}_{M_{i_p^{*_p}}}(M)|, 1 \leq p \leq k\} \times |M|}{|\Delta|}$.

The originality of seasonal gradual patterns as opposed to classical gradual patterns is that they allow to also discover seasonality inside the data in terms of graduality.

After reducing the frequent seasonal gradual patterns mining problem from numerical database to a maximal motif mining problem as illustrated by Table III, we use *MaxMotif* algorithm proposed by [23] on each gradual tuple motif sequence to extract the frequent gradual tuple motifs for given a minimal support threshold. *MaxMotif* algorithm is a polynomial space and polynomial delay algorithm for maximal pattern discovery of the class of motifs with wildcard or joker symbol. It is considered as the most effective specialized approach for enumerating motifs in a sequence. In our approach, we only consider patterns with solid characters, without wildcard symbol. The wildcard symbol for our approach is the character introduced to build gradual tuple motif sequences from gradual tuple motifs as given by definition 6 (we consider the character 'o' as wildcard symbol for our study). The complexity of our proposed approach depends on the complexity of *MaxMotif* algorithm.

Lemma 1: Let Δ be a temporal data sequence over a attribute set $\mathcal{I}$, and M_{i^*} $(i \in \mathcal{I})$ be a gradual tuple motif sequence extracted from Δ. Then $|M_{i^*}| \leq |\Delta|$.

Proof 1: Trivial by using definitions 3 and 7.

V. Experimental results

The major interest of seasonal gradual patterns is that they are well adapted to capture some common co-variations repeated with identical periods on attributes in the ordered data set. One such kind of data that receives a lot of attention nowadays is temporal data, i.e. data produced with a temporal order on the objects, often in e-commerce domain. In order to illustrate the proposed method and show its interest, an experiment study has been conducted on a real data set of customer purchases taking from *Brazilian E-Commerce Public Data set*[1]. This data set contains purchases 99441 transactions of customer purchases on 19 attributes, with an attribute order date on which transactions are ordered. Attribute order date contains different values (different days) which will represent items of the tuple sequences database and other numerical attributes. For our experiments, we retrieve the order days from the order date attribute and consider them as temporal variables $(d_1, d_2, \ldots, d_m)$. The used data set contains 0.56% of missing data, we removed all transactions with missing data and obtained a data set with 99000 transactions.

The experiments are carried out on a 2.8GHz Intel Core i7 CPU, 32GB memory with 8 cores. We focus on the variation of the number of frequent closed seasonal gradual patterns according to the minimum support threshold ($MinSupp$) value, and the computation time required for discovering these seasonal gradual patterns We also show interesting knowledge brought by such patterns in the e-commerce domain. For extracting of maximal motifs in the obtained tuple sequences database, we use *MaxMotif* [23] algorithm which is a linear time algorithm for maximal motif mining. This algorithm is extremely efficient as it can handle huge sequences of over to 10 million length. According to the lemme 1, for the data set used, the longest gradual tuple motif sequence has a length less than or equal to 99000 (the number of transactions).

Figure 1 focus on the variation of the number of frequent closed seasonal gradual patterns according to the minimum support threshold. This figure shows an decrease of number of patterns when the minimal support increases, the number of extracted patterns is even less than 100 for a support threshold less than 0.25 which is easily exploitable by the user.

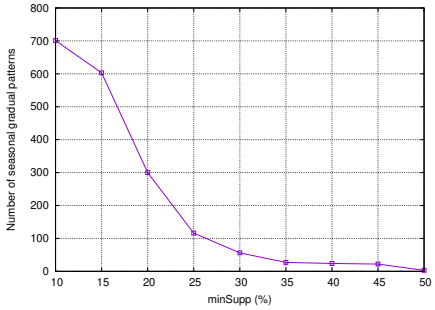

Fig. 1. **Number of seasonal gradual patterns with *minSupp* variation.**

Figure 2 shows the computation time evolution taken by our approach for discovering seasonal gradual patterns according to the minimal support. We observe an decrease of computation time when the support threshold decreases.

Figure 1 and 2 indicate that the number of seasonal gradual patterns is usually small and does not require much time for their extraction, which facilitates their practical exploitation. Table IV shows some interesting seasonal gradual patterns extracted from a e-commerce data set using our approach. Gradual pattern number 4 states that *the higher the price, the lower is the freight value and the higher is the payment value frequently on the temporal sequence* $\langle d_{21}d_{22}d_{23}d_{24}d_{25}\rangle$. This trend of co-variation between the price of products and freight value is also revealed in pattern number 6 with another attribute on another period $\langle d_{18}d_{19}d_{20}d_{21}\rangle$. These patterns could be useful to recommend and to manage business strategies.

[1]https://www.kaggle.com/anshumoudgil/olist-a-brazilian-e-commerce/report

TABLE IV
SOME FREQUENT CLOSED SEASONAL GRADUAL PATTERNS

No.	Seasonal_gradual_patterns	*Support*
1	$\{age^{(\uparrow, d_1 d_2 d_3)}, payment\_installments^{(\uparrow, d_1 d_2 d_3)}\}$	0.59
2	$\{age^{(\uparrow, d_7 d_8 d_9 d_{10} d_{11})}, payment\_value^{(\downarrow, d_7 d_8 d_9 d_{10} d_{11})}\}$	0.23
3	$\{product\_photos\_qty^{(\uparrow, d_2 d_3 d_4 d_5)}, payment\_value^{(\downarrow, d_2 d_3 d_4 d_5)}\}$	0.3
4	$\{price^{(\uparrow, d_{21} d_{22} d_{23} d_{24} d_{25})}, freight\_value^{(\downarrow, d_{21} d_{22} d_{23} d_{24} d_{25})}, payment\_value^{(\uparrow, d_{21} d_{22} d_{23} d_{24} d_{25})}\}$	0.46
5	$\{product\_weight\_g^{(\uparrow, d_{13} d_{14} d_{15})}, freight\_value^{(\downarrow, d_{13} d_{14} d_{15})}, price^{(\uparrow, d_{13} d_{14} d_{15})}\}$	0.33
6	$\{Delivery\_delay^{(\uparrow, d_{18} d_{19} d_{20} d_{21})}, freight\_value^{(\downarrow, d_{18} d_{19} d_{20} d_{21})}, price^{(\uparrow, d_{18} d_{19} d_{20} d_{21})}\}$	0.47

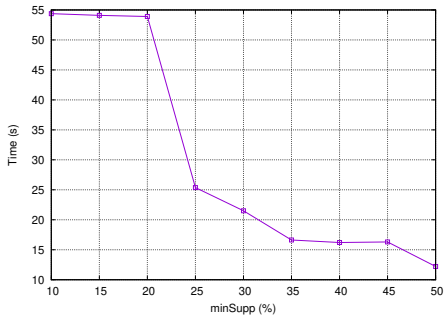

Fig. 2. **Evolution of the computation time for discovering seasonal gradual patterns vs the variation of the *minSupp* value.**

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the seasonal gradual pattern concept which enables us to extract seasonal correlations between attributes from a graduality point of view in a temporal data sequence. The proposed approach formulate in the temporal context, a seasonal gradual patterns mining problem as a problem of finding repeated patterns (frequent pattern) in a sequence with wildcards and exploit existing algorithms for enumeration of maximal motifs in a sequence. We also propose a definition of the associated support measure at a seasonal gradual pattern to efficiently mine frequent patterns in temporal data sequence context. The experimental evaluation on the e-commerce real world data shows the feasibility of our proposed approach and its practical interest for the e-commerce domain.

Future directions aim firstly to enrich the experimental study of the proposed method, to check its applicability to other temporal data sets, data on the flow of product stocks for example. Seasonal gradual patterns extracted from such data set will allow to data experts to detect seasonal co-variations between quantity of products for better management of the supply chain. Another work is to define the approaches for extracting seasonal gradual patterns whose the seasonality are not constituted of consecutive periods.

REFERENCES

[1] S. Ramakrishnan and A. Rakesh, "Mining quantitative association rules in large relational tables," *SIGMOD Rec.*, vol. 25, no. 2, pp. 1–12, 1996.

[2] Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules," in *SIGKDD*, 1999, pp. 261–270.

[3] A. Salleb-Aouissi, C. Vrain, and C. Nortet, "Quantminer: A genetic algorithm for mining quantitative association rules," in *IJCAI*, 2007, pp. 1035–1040.

[4] M. Kaytoue, S. O. Kuznetsov, and A. Napoli, "Revisiting numerical pattern mining with formal concept analysis," in *IJCAI*, 2011, pp. 1342–1347.

[5] L. Di-Jorio, A. Laurent, and M. Teisseire, "Mining frequent gradual itemsets from large databases," in *IDA*, 2009, pp. 297–308.

[6] A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders," *BMC Med*, vol. 17, pp. 133–137, 2013.

[7] E. Hüllermeier, "Association rules for expressing gradual dependencies," in *PKDD*, 2002, pp. 200–211.

[8] F. Berzal, J. C. Cubero, D. Sánchez, M. A. V. Miranda, and J. Serrano, "An alternative approach to discover gradual dependencies," *IJUFKS*, vol. 15, no. 5, pp. 559–570, 2007.

[9] A. Oudni, M. Lesot, and M. Rifqi, "Processing contradiction in gradual itemset extraction," in *FUZZ-IEEE*, 2013, pp. 1–8.

[10] B. Négrevergne, A. Termier, M. Rousset, and J. Méhaut, "Para miner: a generic pattern mining algorithm for multi-core architectures," *DMKD*, vol. 28, no. 3, pp. 593–633, 2014.

[11] T. D. T. Do, A. Termier, A. Laurent, B. Négrevergne, B. O. Tehrani, and S. Amer-Yahia, "PGLCM: efficient parallel mining of closed frequent gradual itemsets," *KAIS*, vol. 43, no. 3, pp. 497–527, 2015.

[12] J. Nin, A. Laurent, and P. Poncelet, "Speed up gradual rule mining from stream data! A b-tree and owa-based approach," *J. Intell. Inf. Syst.*, vol. 35, no. 3, pp. 447–463, 2010.

[13] R. R. Yager, "Families of owa operators," *Fuzzy Sets and Systems*, vol. 59, no. 1, pp. 125–148, 1993.

[14] ——, "On ordered weighted averaging aggregation operators in multi-criteria decisionmaking," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[15] N. Phan, D. Ienco, D. Malerba, P. Poncelet, and M. Teisseire, "Mining multi-relational gradual patterns," in *SDM*, 2015, pp. 846–854.

[16] S. Ayouni, S. B. Yahia, A. Laurent, and P. Poncelet, "Fuzzy gradual patterns: What fuzzy modality for what result?" in *SoCPaR*, 2010, pp. 224–230.

[17] J. Lonlac, Y. Miras, A. Beauger, V. Mazenod, J.-L. Peiry, and E. Mephu, "An approach for extracting frequent (closed) gradual patterns under temporal constraint," in *FUZZ-IEEE*, 2018, pp. 878–885.

[18] T. Ngo, V. Georgescu, A. Laurent, T. Libourel, and G. Mercier, "Mining spatial gradual patterns: Application to measurement of potentially avoidable hospitalizations," in *SOFSEM*, 2018, pp. 596–608.

[19] D. Owuor, A. Laurent, and J. Orero, "Mining fuzzy-temporal gradual patterns," in *FUZZ-IEEE*, 2019, pp. 1–6.

[20] S. Jabbour, J. Lonlac, and L. Saïs, "Mining gradual itemsets using sequential pattern mining," in *FUZZ-IEEE*, 2019, pp. 138–143.

[21] D. Dubois and H. Prade, "Gradual inference rules in approximate reasoning," *Inf. Sci.*, vol. 61, no. 1-2, pp. 103–122, 1992.

[22] T. Calders, B. Goethals, and S. Jaroszewicz, "Mining rank-correlated sets of numerical attributes," in *KDD*, 2006, pp. 96–105.

[23] H. Arimura and T. Uno, "An efficient polynomial space and polynomial delay algorithm for enumeration of maximal motifs in a sequence," *J. Comb. Optim.*, vol. 13, no. 3, pp. 243–262, 2007.

[24] L. Parida, I. Rigoutsos, and D. E. Platt, "An output-sensitive flexible pattern discovery algorithm," in *Combinatorial Pattern Matching CPM*, 2001, pp. 131–142.

[25] N. Pisanti, M. Crochemore, R. Grossi, and M. Sagot, "Bases of motifs for generating repeated patterns with wild cards," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 2, no. 1, pp. 40–50, 2005.

[26] E. Coquery, S. Jabbour, and L. Sais, "A constraint programming approach for enumerating motifs in a sequence," in *ICDMW*, 2011, pp. 1091–1097.