SYSTEM FOR UNCOVERING HIDDEN KNOWLEDGE IN REAL TIME FOR THE ANALYSIS OF ENVIRONMENTAL AND AGRICULTURAL PROCESSES

SISTEMA DE OBTENCIÓN DE CONOCIMIENTO OCULTO EN TIEMPO REAL PARA ANÁLISIS DE PROCESOS MEDIOAMBIENTALES Y AGRÍCOLAS

Francisco Javier Martínez-de-Pisón Ascacíbar

Alpha V. Pernía Espinoza

Roberto Fernández Martínez

Rubén Escribano García

Grupo EDMANS. (www.mineriadatos.com). Universidad de La Rioja. España

Pablo Guillén Rondón

Dante Conti Guillén

Universidad de los Andes. Venezuela

Abstract

The use of data mining techniques in systems that capture information in real time is a field with numerous possibilities that has so far been scarcely exploited. This paper presents the results of the Spanish project CONOSER (DPI2006-03060) on the adaptation of classical data mining algorithms for the real time search for association rules using time series of environmental, industrial and agricultural processes. The aim is to develop systems capable of revealing concealed and useful data to the expert in the form of rules, and that this knowledge can be continuously updated. The system developed captures information on a regular basis, pre-processes it, obtains significant patterns and refreshes the database of frequent items. Using this database, a search is made for those rules that fulfil a series of minimum conditions in terms of support and accuracy, updating the rules and selecting those that may be of use to the expert.

***Keywords**: association rules, analysis of time series, data mining.*

Resumen

La utilización de técnicas de minería de datos en sistemas que capturan la información en tiempo real es un campo poco explotado hoy en día y con amplias posibilidades. En este trabajo, se muestran los resultados del proyecto nacional CONOSER (DPI2006-03060) en la adaptación de algoritmos de minería de datos clásicos para la búsqueda, en tiempo real, de reglas de asociación a partir de series temporales de procesos medio ambientales, industriales o agrícolas. El objetivo, es el desarrollo de sistemas capaces de mostrar conocimiento oculto y útil, en forma de reglas, al experto y que este conocimiento pueda ser continuamente actualizado. El sistema desarrollado captura la información periódicamente, la preprocesa, obtiene patrones significativos y actualiza la base de datos de ítems frecuentes. A partir de esta base de datos, se busca las reglas que cumplan una serie de condiciones mínimas de soporte y precisión, actualizando las reglas y seleccionando aquellas que puedan ser útiles para el experto.

***Palabras clave**: reglas de asociación, análisis de series temporales, minería de datos.*

1. Introduction

Thanks to the reduction in costs and the improvement in Information and Communications Technologies (ICTs), a great deal of information can be obtained on the physicochemical and/or biological variables that have a bearing on environmental and agricultural processes.

The use of wireless sensor networks, together with weather and other kinds of databases, renders it possible to seek concealed information that provides an understanding of the process to be studied or allows developing more efficient predictive models.

The enormous quantity of data stored exceeds the human capacity for processing them without the aid of suitable tools. Multivariate Time Series Databases (TSDBs) are currently a highly valued focus of research for obtaining important and valuable information, which may be extracted by means of Data Mining (DM).

Within the field of Temporal Data Mining (TDM), considerable interest has been shown in recent years in the search for association rules in time series. This interest is centred on obtaining rules that establish relationships between patterns of sundry variables (temperatures, pressures, biological parameters, etc.) within a determined timeframe and with different time intervals.

Tools are currently being designed that, through data mining, allow an analysis to be made of huge databases of recorded information in order to understand environmental, industrial and agricultural processes. One of the goals of the EDMANS research team is to design algorithms and tools that can be used to obtain rules on the basis of the past records of these processes that can help in the decision-making process and in their improvement.

In this article, a type of methodology is considered for seeking association rules that allow for extracting hidden knowledge from databases comprising multivariate time series. The basic idea involves finding, amongst the time series, interrelations between patterns that are frequently repeated and depict these relationships in a manner that is readily understood by an expert.

For example, by analysing the time series corresponding to a certain process, as shown in Figure 1, we could deduce the following rule: "*IF VAR1 rises linearly and VAR2 remains below a given level X, THEN this leads to a drop VAR3*". This previously unknown rule would enable an expert to adopt measures to avoid this drop of VAR3. Of course, it would only be of interest if it was repeated a certain number of times.

It is well known that the human brain's ability to segment and extract visual patterns is much superior to any present system of artificial vision. Likewise, the brain is capable of identifying sound, tastes aromas or textures.

The analyst who seeks to extract some useful knowledge in order to design improvement strategies for a process uses this skill to visually discover repetitive patterns and their interrelationships over time. To do so, the expert uses time graphs like the one shown in Figure 1.

It often happens that when we hear an expert discuss the behaviour of a time series, we hear expressions such as "this segment grows linearly" or "this segment decreases exponentially", which clearly reflect the manner in which human beings locally describe a time series. This type of visual segmentation is performed by dividing the time series into sub-series or segments whose shape or appearance is similar to familiar patterns (lines, rising or falling curves, etc.) just like the way in which the brain describes any new object it encounters.

Once the characteristic segments have been detected, an essential requirement for establishing a possible relationship is the presence of a series of repeated situations in which the same patterns always appear. In the case of Figure 1, VAR1 should rise linearly, VAR2 should remain below a certain level and VAR3 should drop, with this occurring on a significant number of occasions within an appropriate time window.

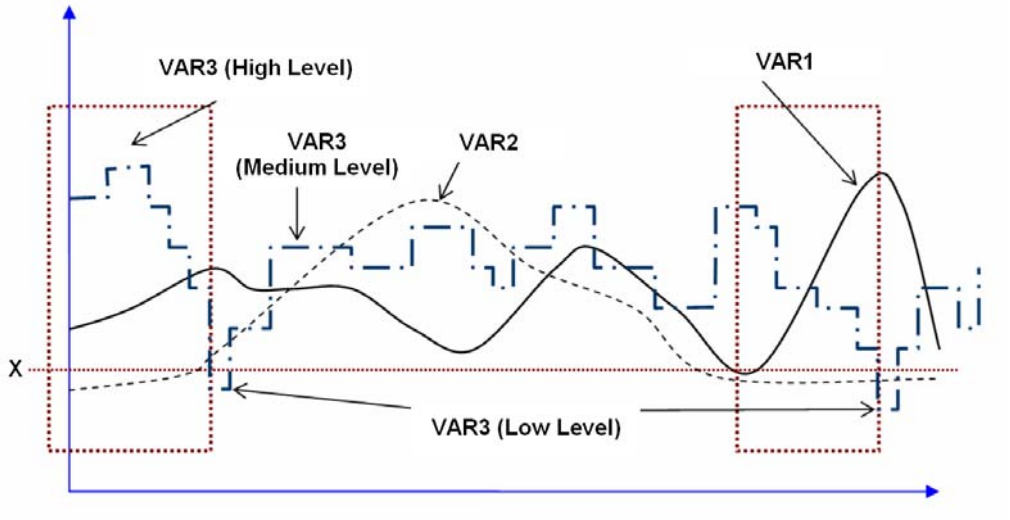

Figure 1. Detecting time relationships between the time series.

2. Association Rules Mining from Time Series

Various authors (Lu et al., 1998; Li et al., 2005; Lee et al., 2007) agree that the classic approach of association rules mining (transactional or intra-transactional approach) widely referenced and studied in the specialist and scientific literature, lacks in great measure of a prediction power that is increasingly required in data interpretation systems, expert systems and decision-making support systems. Association rules (inter-transactional approach) are seen as a way of overcoming this restriction. This approach opens up a whole range of possibilities for studies applied to temporal databases and time series which currently constitute a novel field with great potential in the short to medium term in KDD. A simple example is as follows: take a database of shares and share prices on a stock market, expressed in time series. Within this database, association rules of the following type can be established: *R*(1), If the price of IBM and SUN shares rises, then there is an 80% probability that the price of Microsoft shares will also increase on the same day; and *R*(2), If IBM and SUN share prices increase today, then there is an 80% probability that Microsoft's share price will increase today and also in four days. Although Rule *R(1)* (of intra-transaction type) shows a relationship between the different share prices (during the same day or same transaction), a relationship of type *R*(2) is of much greater interest to a stockbroker, given its greater capacity for prediction or capacity for exploitation. The basic difference between *R*(1) and *R*(2) lies in the presence of an extra dimensionality (time) within Rule *R*(2).

The classic or intra-transactional approach reveals rules such as those exemplified in R(1), i.e., association rules between items in a database, taking into account the same transaction. Something substantially different occurs in R(2): association rules are obtained between field values taken from different transaction records. This concept, which adds a new dimension, is known as inter-transaction association rules mining.

2.1 Association Rules Mining (Transaction Approach). Background

Association rules were introduced by Agrawal et al. (1993) with the classic problem of the shopping basket. These references indicate that there are hidden relations between the items acquired in a transactional database. These relations could explain new behaviour in shopping habits among customers in a supermarket which had not previously been considered or perceived.

These relations, called association rules, are sentences of type:

$$X \rightarrow Y \tag{1}$$

where *X*, *Y* (*X* implies *Y*, *X* called the antecedent and *Y* the *consequent*) are sets of frequent items in a given database such that:

$$X \cap Y = \Phi \tag{2}$$

The *support* of Rule $X \rightarrow Y$ represents the percentage of transactions in the database that contain both *X* and *Y*, i.e. $P(X \cup Y)$. The *confidence* of the rule is the percentage of transactions in the database containing *X* that also contain *Y*, i.e.:

$$P\left(\frac{X}{Y}\right) \tag{3}$$

Association rules mining is therefore used to find the relations between sets of items in a database, in such a way that the support and confidence of these rules meets the minimum levels of support and confidence previously established by the analyst.

The basic problem of association rules can therefore be divided into two sub-problems: The first consists of finding sets of frequent items that exceed the limit or level of support established by the analyst and the second, finding the association rules per se using the frequent items obtained from the first subproblem with a minimum level of confidence that enables the resulting rules of the iterative search process to be distinguished.

2.2 The inter-transactional approach

Before outlining the problem of inter-transaction association rules we should consider some basic definitions and the context in which it lies. This reference emphasises the need to differentiate inter-transaction association rules mining from the sequential pattern mining approach by establishing a point of divergence between the two techniques. Within the search for sequential patterns, each customer's transactions over time (temporal dimension) are treated as if they were a single transaction. This means that the rules or patterns discovered are in nature of intra-transaction type in that each sequence is indeed treated as a single transaction; and the mining process will focus on finding similarities between these sequences. The inter-transactional approach is opposed to this principle since it seeks to find values/relations among different transaction records in the database. Although some authors cite inter-transaction association rules as being extensions of the discovery of sequential patterns and episodes, the trend, as described above, is to establish the fundamental difference from the perspective of processing in the transaction.

However, according to some authors (Lu et al., 1998; Li et al., 2005) an initial point of convergence could also be established between the two techniques. The work reported by Mannila et al. (1997) on episode mining is similar to the dimensional concept (time) in inter-transaction association rules.

An episode is a sequence of events and the association rules between the episodes take the form:

$$P(V) \rightarrow Q(W) \tag{4}$$

where *P* and *Q* are a sequence of events and *V*, *W* are time limits. Nonetheless, the concept of transaction within the sequence of events is not clear. Although sub-sequences are formed using time limits and the episodes that occur frequently in those sub-sequences are discovered with extensions of the *Apriori* algorithm, the procedure appears to indicate that the results obtained are intra-transactional as opposed to inter-transactional. Similarly, Bettini et al. (1998) propose the use of structures of events to discover temporal relationships between events in a time sequence (time series). This structure consists of a number of variables that represent temporal events and constraints between these variables, which are later subjected to a search for association rules. Here once again, there is no clear concept of transaction, and therefore, these advances partly disagree with the inter-transactional approach.

3. Association Rules in Environmental Engineering Sciences

In (Feng et al., 2001) an application of the association rules for predicting the weather in an observatory in Hong Kong is presented. The study establishes rules in weather variables - wind direction, wind speed, temperature, humidity and atmospheric pressure - to predict the state of the weather (Rain). Another notable feature of this study is the dimensionality, since it not only includes the time attribute, but also the locations of meteorological stations in different parts of Hong Kong.

Likewise, (Harms et al., 2003) analyse the phenomenon of drought in the state of Nebraska, USA, relating it to oceanic and atmospheric variables and using historical records compiled from 1950 to 1999. Harms (Harms, 2004) extrapolates the previous study with the addition of support systems for geospatial decisions on managing environmental or meteorological risks related to drought phenomena. Two algorithms of association rules for time series are developed together with two methods of interpolation. The purpose is to discover association rules between environmental variables and drought events. The association rules are determined by an episode labelled as an antecedent and an episode labelled as a consequent, where an episode is a sequence of event(s). It uses the search algorithms of association rules of REAR and MOWCATL (Harms et al., 2003).

Also, at the end of the 2007, Huang et al. (2007) present an application referred to environmental data with oceanic variables. The paper shows an efficient technique for analyzing ARGO ocean data comprising time series of salinity/temperature measurements where informative salinity/temperature patterns are extracted by using inter-transaction rules. By this way, they found important information of the associated salinity/temperature variations among different locations and time intervals. A quantitative inter-transaction association rules mining algorithm is proposed. The FITI and the PrefixSpan algorithms are adopted to maximize the mining efficiency.

4. Proposed Methodology

Following a study of existing literature and a thorough analysis of current applications, the initial methodology has been partially modified for the phases of pre-processing and segmentation of time series in order to obtain useful patterns. Searching for repeated

sequences of patterns was made using a pre-established time window. Finally, a classical inter-transactional association rule's algorithm was used.

An extensive study was conducted initially on current techniques in pre-processing, segmentation and the search for frequent patterns. The most significant conclusion this study reached was that no combination of the huge amount of existing techniques is 100% effective for the pre-processing and automatic segmentation of the time series from industrial processes (Martínez-de-Pisón et al., 2005). This is due to the total heterogeneity that exists between the various types of time series from an industrial process (temperature, pressure, connection or disconnection of a motor, etc.), which means that the expert has to work in an iterative and interactive manner with several pre-processing and segmentation algorithms for each kind of time series to be segmented.

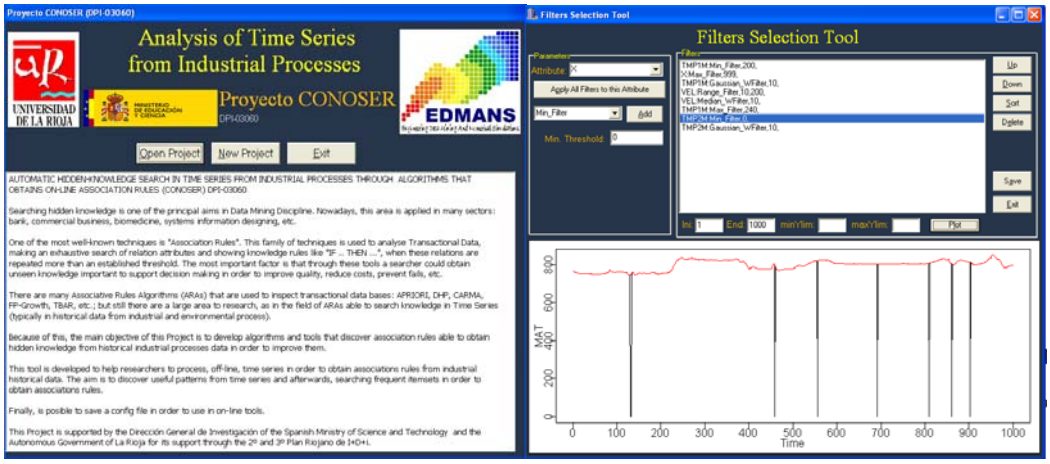

Figure 2. Main screen of CONOTOOL software and the "Filter Selection" section.

Accordingly, the focus on the methodology in all these phases was directed towards the development of several pre-processing and segmentation techniques that were of straightforward application by the expert analyst of the industrial process. All the pre-processing and segmentation algorithms of time series are being implemented for the free statistical analysis program R (R Development Core Team, 2008) within a library called *KDSeries* which will be delivered to the community of R users under GPL licence.

Furthermore, given the enormous effort in trial and error that is required during these stages, the decision was taken to develop a new visual tool, called *CONOTOOL* (Figure 2), which will make these tasks easier. This tool allows for a more intuitive and speedier use of all the functions in the *KDSeries* library. The latest version available, of both the *KDSeries* library and *CONOTOOL*, can be downloaded from the CONOSER Project website: http://api.unirioja.es/conoser.

The proposed methodology comprises the following stages for a time period (one week, one month, etc.):

1. Filtering of each time series to eliminate noise and obtain its basic form.
2. Obtaining important minimum and maximum points.
3. Extraction of characteristics sub-patterns ("incremental (INC)", "decremental (DEC)", "horizontal (HOR)" or "according to a threshold (UMB)") (Figure 3, left).
4. Grouping of the "INC", "DEC", "HOR" and "UMB" sub-patterns into more complex patterns according to the specifications made by the expert (Figure 3, right).

5. Searching for sequences of patterns using a sliding window and creating data streams (see Figure 4).
6. Mining frequent itemsets over data streams online and incrementally within a sliding window.
7. Extracting on-line association rules over data streams.

4.1 Filtering of each time series to eliminate noise and obtain its basic form

Some of the techniques that are already being implemented in the *KDSeries* library and in the visual tool *CONOTOOL* allow for the performance of sliding-window filters under Kernel (Gaussian, rectangular, maximum, minimum, median, etc.), eliminating according to a threshold (according to a minimum or maximum value or a range), based on the Fast Fourier Transform (FFT) to filter high o low frequencies harmonics, linear approach to series, piecewise constant approximation, constant approach to series, aggregate approach to series or adaptive constant approach to series (Keogh et al., 2004; Hetland, 2004). Regarding the extraction techniques for characteristics, it is worth noting that the techniques have been realigned to make them more useful in the iteration with users, as automatic extraction has not given very good results, although they remain available in the KDSeries library.

To facilitate the job of analysts in this critical process, an iterative system was implemented in CONOTOOL which enables numerous filters to be applied sequentially to each time series. Researchers can thus quickly adjust the parameters of each filter until the basic form of the time series is obtained.

Using the filter tool (Figure 2), we can allocate different filters to each time series. We can view the effect a single filter has ("*Plot*" button) on the time series or view the impact of a series of filters applied to it ("*Apply all filters to this Attribute*" button). The original series appears in black and the filtered series in red.

The order of application of the filters is chosen according to the order in which they appear on the screen (top-down). Clicking on the buttons up, down, delete, etc. allows for conveniently reorganising the filter bank.

The filters allow for:

- Deleting the data below, above and between a range or outside a range of threshold values: Min_Filter(), Max_Filter(), Range_Filter(), InvRange_Filter(), respectively.
- Using a sliding-window filter type that is Gaussian, median, mean, minimum or maximum: Gauss_Filter(), Median_Filter(), Mean_Filter(), Min_Filter() and Max_Filter(), respectively.
- Using the Fast Fourier Transform (FFT), applying a rectangular or Gaussian window: FFT_Filter_Mean() FFT_Filter_Gauss().

4.2 Obtaining important minimum and maximum points

Once the signal has been filtered, important minimum and maximum points are obtained for each time series in order to identify increasing, horizontal and decreasing patterns (see Figure 3). The intuitive idea is to discard minor fluctuations and keep major minima and maxima (Fink and Pratt, 2003).

We can considerer a point a_m of a series $a_1,...,a_n$ as a *major minimum* if there are indices i and j such that:

$$a_m \text{ is the minimum among } a_i,...,a_j \text{ and } \frac{a_i}{a_m} \geq R \text{ and } \frac{a_j}{a_m} \geq R \ (where,\ i \leq m \leq j) \qquad (5)$$

and a_m as an *major maximum* if:

$$a_m \text{ is the maximum among } a_i,...,a_j \text{ and } \frac{a_m}{a_i} \geq R \text{ and } \frac{a_m}{a_j} \geq R \ (where,\ i \leq m \leq j) \qquad (6)$$

where *R* is a compression rate always greater than one.

4.3 Extraction of characteristics sub-patterns

The algorithm caters for the inclusion of a search bench for different subpatterns in each time series according to the height, width and type of curve in order to obtain “incremental (INC)”, “decremental (DEC)”, “horizontal (HOR)” or “according to a threshold (UMB)” steps as per values established previously for each time series.

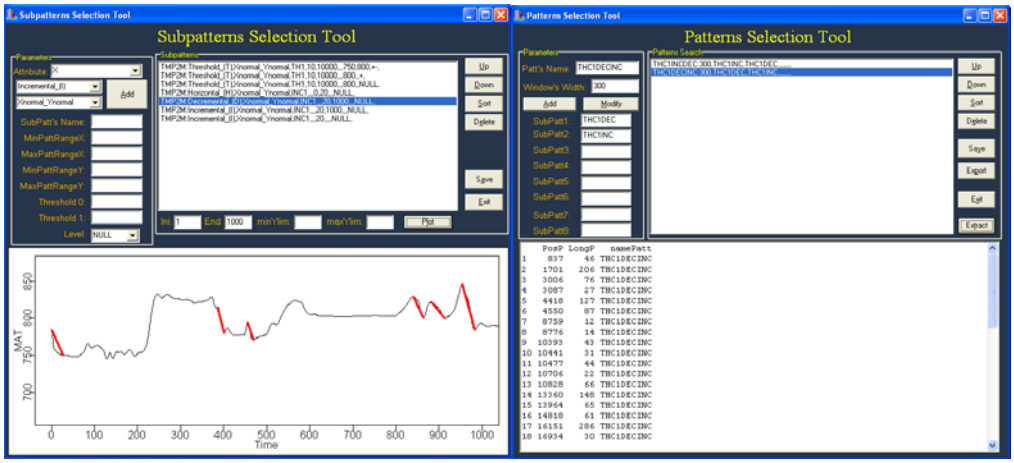

Figure 3. Subpatterns and Patterns Selection Tool.

Definition 4.3.1

An *incremental subpattern (INC)* is found if two consecutive major points a_k and a_l of a series $a_1,...,a_n$ (where $k < l$) satisfy the following:

$$\begin{aligned} &a_k \text{ is a minimum, and } a_l \text{ is a maximum, and } (l-k) \geq w_1, \text{ and} \\ &(l-k) \leq w_2, \text{ and } a_l - a_k \geq h_1, \text{ and } a_l - a_k \leq h_2 \end{aligned} \qquad (7)$$

where $\{w_1, w_2\}$ is the range formed by a vector of two values of X within which the subpattern is to be contained and $\{h_1, h_2\}$ is the range formed by a vector of two values of Y within which the subpattern is to be contained.

Definition 4.3.2

A *decremental subpattern (DEC)* is found if two consecutive major points a_k and a_l of a series $a_1,...,a_n$ (where $k < l$) satisfy the following:

$$\begin{aligned} &a_k \text{ is a maximum, and } a_l \text{ is a minimum, and } (l-k) \geq w_1, \text{ and} \\ &(l-k) \leq w_2, \text{ and } a_k - a_l \geq h_1, \text{ and } a_k - a_l \leq h_2 \end{aligned} \quad (8)$$

where $[w_1, w_2]$ is the range formed by a vector of two values of X within which the subpattern is to be contained and $[h_1, h_2]$ is the range formed by a vector of two values of Y within which the subpattern is to be contained.

Definition 4.3.3

An *horizontal subpattern (HOR)* is found if two consecutive major points a_k and a_l of a series $a_1,...,a_n$ (where $k < l$) satisfy the following:

$$(l-k) \geq w_1, \text{ and } (l-k) \leq w_2, \text{ and } |a_k - a_l| \leq h_2 \quad (9)$$

where $[w_1, w_2]$ is the range formed by a vector of two values of X within which the subpattern is to be contained and $[0, h_2]$ is the range formed by a vector of two values of Y within which the subpattern is to be contained.

Definition 4.3.4

An "*over a threshold" subpattern (UMB)* is found if two consecutive major points a_k and a_l of a series $a_1,...,a_n$ (where $k < l$) satisfy the following:

$$(l-k) \geq w_1, \text{ and } (l-k) \leq w_2, \text{ and } a_k > t, \text{ and } a_l > t \quad (10)$$

where $[w_1, w_2]$ is the range formed by a vector of two values of *X* within which the subpattern is to be contained and t is a threshold value.

Definition 4.3.5

A "*below a threshold" subpattern (UMB)* is found if two consecutive major points a_k and a_l of a series $a_1,...,a_n$ (where $k < l$) satisfy the following:

$$(l-k) \geq w_1, \text{ and } (l-k) \leq w_2, \text{ and } a_k < t, \text{ and } a_l < t \quad (11)$$

where $[w_1, w_2]$ is the range formed by a vector of two values of *X* within which the subpattern is to be contained and t is a threshold value.

Regarding these tasks, techniques were initially developed that enabled frequent patterns to be extracted automatically. The problem is that, given the vast heterogeneity of the time series, it was observed that the results were not fully satisfactory. Although these techniques continue to be available in the *KDSeries* library, new techniques have been developed that can be wholly configured by the expert.

4.4 Grouping subpatterns into more complex patterns

Once we have obtained the subpatterns for each time series, we then use the pattern extraction tool (Figure 3, right). It defines the sequence of patterns to be found within a window defined by the user. If that sequence is found, it creates a new pattern with the name specified by the user. It is important to note that wild cards can be used for complex searches like the *grep()* instruction in *Perl* language. For example, with the following search:

Window's width=300, SubPatt1="TEMPINC", SubPatt2="*", SubPatt3="PRESHIGH"*

The program will search, within a time window of 300, for those sub-sequences whose first subpattern begins with "*TEMPINC*", the second subpattern can be anything and the third subpattern is "*PRESHIGH*". The *CONOTOOL* software stores the position of the patterns found and allocates them the name specified by the user.

4.5 Searching for sequences of patterns using a sliding window to create data streams

In each new pattern founded, a search is made for sequences of important patterns (transactions) that are repeated amongst the various time series according to a pre-established time sliding window. Transaction data stream is formed by this continuous sequence of transactions (see Figure 4).

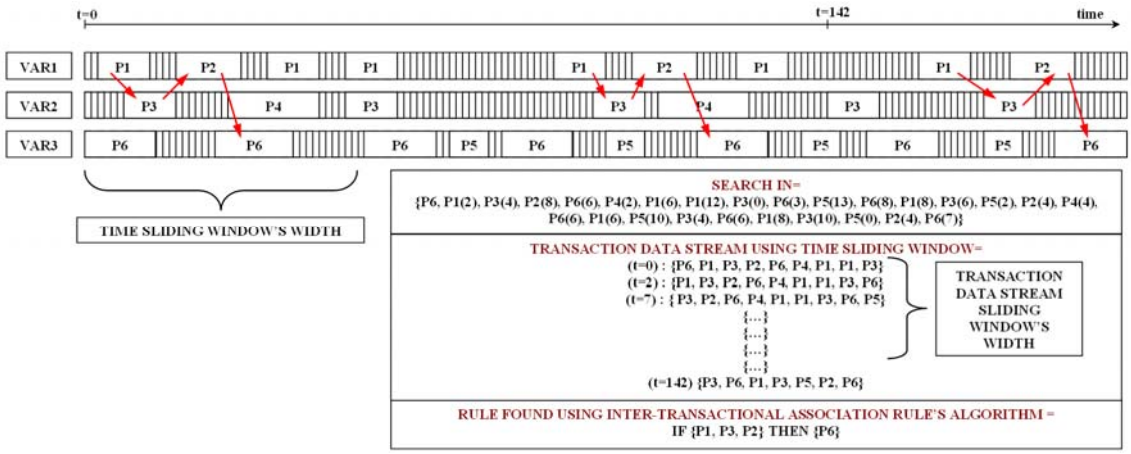

Figure 4. Searching for sequences of patterns using a sliding window.

4.6 Mining frequent itemsets over data streams online and incrementally within a sliding window

Using a transaction data stream, the next objective is to obtain the frequent itemsets that exceed a pre-established support.

There are numerous algorithms for extracting frequent items from data streams, being classified into three main groups. An extensive analysis is made in (Cheng et al., 2008) of the different algorithms existing, highlighting each one's strengths and weaknesses.

In order to proceed to the real-time and continuous extraction of frequent items from the transaction data stream, use has been made of the algorithm MFI-TransSW (**M**ining **F**requent **I**temsets within a **T**ransaction-sensitive **S**liding **W**indow) created recently by Li and Lee (2009).

This algorithm makes online and incremental mining of frequent itemsets in data streams within a transaction-sensitive sliding window. A transaction-sensitive sliding window is composed of a fixed number of transactions. An effective bit-sequence representation of item is developed to maintain the sliding order of window and the itemsets frequencies.

MFI-TransSW is highly efficient, both in terms of the use of runtime and of memory, for the single-step extraction of items from a transaction data stream. A description is provided in (Li and Lee, 2009) of this and other algorithms, along with a thorough analysis of the state-of-the-art of these techniques, and a comparison is made between MFI-TransSW and the other more widely-used algorithms.

4.7 Extracting on-line association rules over data streams

Finally, extraction is made of the inter-transactional association rules for frequent itemsets (Fig. 4).

Definition 4.7.1

Let, $S= \{s_1, s_2,\ldots, s_u\}$ be a set of time series and $Ti = \{s_1(i),s_2(i),\ldots s_u(i)\}$ $(1\leq i\leq n)$ where T_i *as the* values of set S in time *i* (Dimensionality 1),).

The union set of multiple time series D is then defined as:

$$D=\{T_1,T_2,\ldots,T_n\} . \quad (12)$$

Definition 4.7.2

Let $\Sigma_1 =\{e_1, e_2 , \ldots,e_u \}$ be a set of events. These will be attributes of the time series

$$S.\Sigma = \{e_1(0),\ldots, e_1(w-1), e_2(0),\ldots, e_2(w- 1) , \ldots, e_u(0),\ldots, e_u(w-1) \} \quad (13)$$

as the set of possible extension of Σ_1. Let *w be the sliding window* in D, and taking *s* as a time reference $(1 \leq s \leq n-w+1)$, if e_i *occurs in the time* $s+x$ $(0 \leq x \leq w-1)$, then it is labelled that $e_i(x)$ belongs to T_s.

Definition 4.7.3

An association rule of inter-transaction type in a multiple time series is an implication of the form $X \iint Y$, which satisfies:

$$X\subseteq \Sigma,\ Y\subseteq \Sigma,\ X\cap Y=\Phi \quad (14)$$

$$! \approx e_i(0)\succeq X,\ 1\leq i\leq u \quad (15)$$

$$! \approx e_j(q)\succeq X,\ 1\leq j\leq u,((i=j)\wedge (1\leq q<w-1))\preceq ((i\neq j)\wedge (0\leq q<w-1)) \quad (16)$$

and

$$! \approx e_i(p)\succeq Y,\ 1\leq i\leq u,\ \max(q)<p\leq w-1. \quad (17)$$

Let *n* be the number of transactions; C_{XY} the number of times that $X \nearrow Y$ appears in the database; C_X the number of times that *X* appears in the database, therefore the support and the confidence of the inter-transaction rule will be:

$$Type = C_{XY} / n \quad Confidence = C_{XY} / C_X. \quad (18)$$

Definition 4.7.4

An intra-transaction itemset is a set in which all items are taken from the same transaction T_i, whereas an inter-transaction itemset is a set in which the items are taken from multiple transactions.

For this type of set, the following lemma is proved: Let F be an itemset of inter-transaction type,

$$A_i = \{ e_j \mid 1 < j < u,\ e_j(i) \succ F \} \text{ with } 0 \le i \le (w-1). \quad (19)$$

Then, for any i, 0≤ i ≤ (w-1), Ai must be an itemset of intra-transaction type. This lemma is extremely important since an inter-transaction itemset may be considered as a combination of various itemsets of intra-transaction type. Therefore, the inter-transaction-type rules could be generated from intra-transaction-type rules mining.

Taking the problem defined, we proceed to search for inter-transaction association rules with classical algorithms (Apriori, FP-Growth or Eclat).

5. Conclusions

The aim of this article has been to show a methodology which is based on simple preprocessing, segmentation and the search for hidden knowledge on the basis of time series which can be easily applied to enhance processes.

On a practical level, the conclusion drawn is that by means of iterative and interactive adjustments of the pre-processing and segmentation functions (filters, detection of major maxima and minima, extraction of subpatterns and patterns) experts can obtain significant patterns more reliably than in fully automatic pattern extraction systems. We therefore consider it worthwhile making the effort required to develop tools to facilitate iteration between experts and analysis software in all pre-processing and time series segmentation tasks.

It is worth noting that the use of tools of this kind can be used not only for extracting knowledge from environmental, agricultural o industrial processes but that it can also be extrapolated, with the same degree of success, to other fields: business, marketing, etc.

Acknowledgments

The authors thank the “Dirección General de Investigación” of the Spanish Ministry of Science and Innovation for the financial support of the projects DPI2006-03060, DPI2006-14784, DPI-2006-02454 and DPI2007-61090; and the European Union for the project RFS-PR-06035.

Finally, the authors also thank the Autonomous Government of La Rioja for its support through the 3º Plan Riojano de I+D+i.

References

Agrawal R., Imielinski T. and Swami A. “Database Mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering”, Special Issue on Learning and Discovery in Knowledge-Based Databases, Vol. 5, 1993, pp. 914-925.

Bettini C., Wang X. S. and Jajodia, S., “Mining temporal relationships with multiple granularities in time sequences”, Data Engineering Bulletin, Vol. 21, 1998, pp. 32-38.

Cheng J., Ke Y. and Ng W., “A Survey on Algorithms for Mining Frequent Itemsets over Data Streams”, Knowledge and Information Systems, Vol. 16, 2008, pp. 1-27.

Feng L., Dillon T. and Liu J., “Inter-transactional association rules for multi-dimensional contexts for prediction and their applications to studying meteorological data”, Data & Knowledge Engineering, 37, 2001, pp. 85-115.

Fink E. and Pratt K. B., “Indexing of compressed time series”. In M. Last, A. Kandel, and H. Bunke, editors, Data Mining In Time Series Databases, pp. 43-64. World Scientic, 2004.

Harms S., Tadesse T., Wilhite D., Hayes M. and Goddard S., “Drought Monitoring Using Data Mining Techniques: A case study for Nebraska, USA.”, Natural Hazards, 33, 2004, pp. 137-159.

Harms S., Deogun J. and Tadesse T., “Discovering sequential association rules with constraints and time lags in multiple sequences”. Proceedings of the 13th International Symposium on Foundations of Intelligent Systems, Springer 2002, pp.432-441.

Harms S., “Time Series data mining in a Geospatial Decision Support System”, ACM International Conference Proceedings Series, Vol. 130, 2003, pp. 1-4.

Hetland M. L., “A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences”. Series in Machine Perception and Artificial Intelligence. Vol. 57, 2004, pp. 23-42.

Huang Y., Hsu C. and Wanga S., “Pattern recognition in time series database: A case study on financial database”, Expert Systems with Applications: An International Journal, Vol. 33, 2007, pp. 199-205.

Keogh E., Chu S., Hart D. and Pazzani, “M. Segmenting Time Series: A Survey and Novel Approach”. Series in Machine Perception and Artificial Intelligence. Vol. 57, 2004, pp. 1-21.

Lee A. and Wang, C., “An efficient algorithm for mining frequent inter-transaction patterns”. Information Sciences, Vol. 177, 2007, pp. 3453-3476.

Li Q., Feng L. and Wong, A. “From intra-transaction to generalized inter-transaction: Landscaping multi-dimensional contexts in association rule mining”. Information Sciences, Vol. 172, 2005, pp. 361-395.

Li H-F and Lee S-Y, “Mining frequent itemsets over data streams using efficient window sliding techniques”, Expert Systems with Applications, Vol. 36, 2009, pp. 1466-1477.

Lu H., Han J. and Feng, L., “Stock Movement Prediction and n-dimensional Inter-transaction association rules”. Proceedings of the ACM-SIGMOD workshop on research issues on Data Mining and Knowledge Discovery, 1998, pp. 1-12.

Manilla H., Toivonen H. and Verkamo, A., “Discovery of frequent episodes in event sequences”. Data Mining Knowledge Discovery, Vol. 1, 1997, pp. 259-289.

Martínez-de-Pisón F.J., Pernía A., Castejón M. and González A., “Descubrimiento de Conocimiento Mediante Técnicas de Minería de Datos Aplicadas a Series Temporales para la Optimización de Procesos Industriales”, Proceedings of the IX International Congress of Project Engineering, Málaga 2005, Spain.

R Development Core Team, “R: A language and environment for statistical computing. R Foundation for Statistical Computing”, Vienna, Austria, 2008. URL http://www.R-project.org.

Correspondence:

Dr. Francisco Javier Martínez de Pisón Ascacíbar
Grupo EDMANS. URL: http://www.mineriadatos.com
Área de Proyectos de Ingeniería. Departamento de Ingeniería Mecánica
Edificio Departamental. ETSII de Logroño. C/ Luís de Ulloa, 20, 26004 Logroño (España).
Phone: +34 941 299 232. Fax: + 34 941 299 794
E-mail: fjmartin@unririoja.es.