Analysing Satellite Image Time Series by means of Pattern Mining

François Petitjean, Pierre Gançarski, Florent Masseglia, Germain Forestier

Analysing Satellite Image Time Series by means of Pattern Mining

* François Petitjean[1], Pierre Gançarski[1], Florent Masseglia[2], and Germain Forestier[1]

[1] LSIIT (UMR 7005 CNRS/UdS) – Bd Sébastien Brant – 67412 Illkirch – France
[2] INRIA Sophia Antipolis – 2004 route des lucioles – 06902 Sophia Antipolis – France
fpetitjean@unistra.fr

Abstract. Change detection in satellite image time series is an important domain with various applications in land study. Most previous works proposed to perform this detection by studying two images and analysing their differences. However, those methods do not exploit the whole set of images that is available today and they do not propose a description of the detected changes. We propose a sequential pattern mining approach for these image time series with two important features. First, our proposal allows for the analysis of all the images in the series and each image can be considered from multiple points of view. Second, our technique is specifically designed towards image time series where the changes are not the most frequent patterns that can be discovered. Our experiments show the relevance of our approach and the significance of our patterns.

1 Introduction

As remote sensing has witnessed an important technological progress with high definition images, another progress is taking shape with satellites (*e.g.* $Venμs$, *Sentinel-2*) able to acquire image time series at high frequency (two, three images a week and even more). These Satellite Image Time Series (SITS) are an important source of information for scene (*i.e.* geographic area) analysis. A possible but naive usage of these images would consist in selecting two images from the series and study their differences and the evolutions they reveal. However, changes in a scene might spread over a long time period (urbanization, for instance, lasts for several years and building sites do not have the same start time and end time) or they might cycle (such as crop rotation). Consequently, the number of possible combinations is intractable and cannot be reduced to the analysis of two images. We propose to analyse a scene with satellite images on important time periods (our approach will be tested over 35 images and a period of 20 years).

Our approach combines an adequate transform of satellite images and a targeted sequential pattern mining algorithm [1, 2]. This family of algorithms is

* This work was supported by the CNES (French space agency) and by Thales Alenia Space.

typical of knowledge discovery and allows to discover regular or frequent patterns from a set of records. Here a record will be the values of one pixel (*i.e.* its evolution in time). Let us consider, for instance, a set of 24 satellite images of Dubaï, over a period of 2 years (1 image each month). An expected frequent pattern that would be discovered from such a dataset would probably be "15% of all pixels are typical of a desert, then they have the characteristics of a building site and then the characteristics of buildings". In other words if there exists a large enough set of pixels with the same "behaviour" (*i.e.* these pixels have the same evolution), then this behaviour must be discovered. Let us mention that the pixels' position is not a criteria here (our goal is not to extract pixels because of a shape). Our goal is to extract significant schemas in the evolution of a set of pixels.

Mining sequential patterns from satellite images makes sense, since pixels having the same evolution will be characterized by the same pattern. Once discovered, these specific schemas of evolution will be given to experts for validation. Examples of such schemas can be found in urbanization (like in Dubaï for instance) or in road creation (where the schema would contain "vegetation" followed by "bare soil" followed by "road").

This paper if organized as follows. In Sect. 2 we give an overview of existing works in SITS analysis. Section 3 gives the main definitions of sequential patterns and Sect. 4 describes the preprocessing of SITS for the discovery of such patterns. In Sect. 5 we propose a sequential pattern mining technique devoted to SITS and our results are described in Sect. 6. Eventually, we conclude this paper in Sect. 7.

2 Related Works: SITS Analysis

Change detection in a scene allows the analysis, through observations, of land phenomenon with a broad range of applications such as the study of land-cover or even the mapping of damages following a natural disaster. These changes may be of different types, origins and durations.

In the literature, we find three main families of change detection methods. Bi-temporal analysis, *i.e.*, the study of transitions, can locate and study abrupt changes occurring between two observations. Bi-temporal methods include image differencing [3], image ratioing [4] or change vector analysis (CVA) [5]. A second family of mixed methods, mainly statistical, applies to two or more images. They include methods such as post-classification comparison [6], linear data transformation (PCA and MAF) [7], image regression or interpolation [8] and frequency analysis (*e.g.*, Fourier, wavelet) [9]. Eventually, we find methods designed towards image time series and based on radiometric trajectory analysis [10].

Whatever the type of methods used in order to analyse satellite image time series, there is a gap between the amount of data representing these time series, and the ability of algorithms to analyse them. First, these algorithms are often dedicated to the study of a change in a scene from bi-temporal representation.

Second, even if they can map change areas they are not able to characterize them. As for multi-date methods, their results are usually easy to interpret and do not characterize the change.

Meanwhile, frequent sequential pattern mining [1, 2] is intending to extract patterns of evolution in a series of symbols. These methods allow to identify sets of sequences that had the same underlying evolution. Furthermore, they are able to characterize this evolution, by extracting the pattern shared by this set of sequences.

Extracting frequent sequences from SITS was introduced in [11]. The authors study the advantages of such sequences in two applications: weather and agronomics. However, their proposal allows discovering sequences on series of images where the pixels can have only one value.

Our proposal, as explained in Sect. 4, applies to images where the pixels take values on tuples, each value corresponding to a separate band. This characteristics, along with the large number of images, will have important consequences on the patterns, their relevance and the complexity of their discovery.

3 Mining frequent sequential patterns

Sequential patterns are extracted from large sets of records. These records contain sequences of values that belong to a specific set of symbols, as stated by definition 1 (inspired by the definitions of [1]).

Definition 1. *Let $\mathcal{I} = \{i_1, i_2, ..., i_m\}$, be a set of m values (or* items*). Let $I = \{t_1, t_2, ...t_n\}$, be a subset of $\mathcal{I}$. I is called an* itemset*. A* sequence *s is a non-empty list of itemsets noted $< s_1, s_2, \ldots, s_n >$ where s_j is an itemset. A* data sequence *is a sequence in the dataset being analysed.*

Definition 2 shows the conditions for the inclusion of two sequences. In other words, s_1 is included in s_2 if each itemset of s_1 is included in an itemset of s_2 with the same order. This definition is illustrated by Example 1.

Definition 2. *Let $s_1 =< a_1, a_2, \ldots, a_n >$ and $s_2 =< b_1, b_2, \ldots, b_m >$ be two sequences. s_1 is* included *in s_2 ($s_1 \prec s_2$) if and only if $\exists i_1 < i_2 < \ldots < i_n$ integers, such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}, \ldots$ $a_n \subseteq b_{i_n}$.*

Exemple 1 *The sequence $s1 =<$ (3) (4 5) (8) $>$ is included in the sequence $s2 =<$ (7) (3 8) (9) (4 5 6)(8) $>$ (*i.e.*, $s1 \prec s2$) since* (3) $\subseteq$ (3 8), (4 5) $\subseteq$ (4 5 6) *and* (8) $\subseteq$ (8)*. Meanwhile, the sequence $s3 =<$ (3 8 9) (4 5) $>$ is not included in s_2 since* (3 8 9) *is not included in an itemset of* $s2$*.*

In this paper, the main characteristic for sequential pattern extraction will be their frequency. This notion is based on the number of occurrences of a pattern, compared to the total number of sequences, as stated by definition 3. Eventually, for simplicity in the results, only the longest patterns are kept (C.f. definition 4).

Definition 3. *A data sequence* s_d supports *a sequence* s *(or participates in the support of* s*) if* $s \prec s_d$*. Let* D *be a set of data sequences. The* support *of* s *in* D *is the fraction of data sequences in* D *that support* s*:* $support(s) = |\{s_d \in D / s \prec s_d\}|/|D|$*. Let minSupp be the minimum support value, given by the end-user. A sequence having support higher than minSupp is frequent.*

Definition 4. *Let* F^D *be the set of frequent sequential patterns in* D*. In a set of sequences, a sequence* s *is maximal if* s *is not contained in any other sequence. Let* L^D *be the set of maximal sequences of* F^D*.* L^D *is the set of* maximal frequent sequential patterns *in* D*.*

4 Data preprocessing (SITS)

We want to analyse images from the Kalideos database (the scenes are located in the south-west of France). We have extracted a series of 35 images as illustrated by Fig. 1. These images cover a period of 20 years.

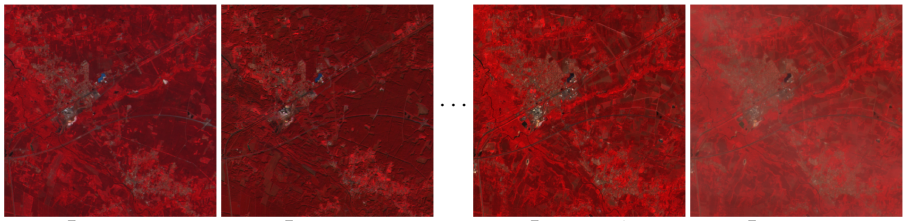


Fig. 1. Extract of the Satellite Image Time Serie of KALIDEOS used. © CNES 2010 – Distribution Spot Image

Since these images were acquired by different sensors, the comparison of radiometric levels of a pixel (x, y) from one image to another calls for corrections. The value of each pixel has to be adjusted. First, we need to make sure that pixel (x, y) in a series cover the very same geographic localization in every image. Then, some corrections are performed by the CNES in order to reduce the impact of atmospheric changes from one picture to another (since two pictures can be separated by several months).

Once the corrections performed, we are provided with 35 images where each pixel takes values on three bands: Near Infra-Red (NIR), Red (R) and Green (G). To these bands, we add a fourth one, corresponding to the Normalized Difference Vegetation Index (NDVI) calculated as follows for a pixel p: $\text{NDVI}\,(p) = \frac{\text{NIR}_{(p)} - \text{R}_{(p)}}{\text{NIR}_{(p)} + \text{R}_{(p)}}$

Then, each sequence is built as the series of tuples (NIR,R,G,NDVI) for each pixel (x, y) in the image series.

Eventually, a discretization step is necessary on the bands' values for a sequential pattern extraction. Actually, this step will lower the total number of items during the mining step. Therefore, on each band, we have applied a K-MEANS algorithm [12] in order to obtain 20 clusters of values. For readability, the cluster numbers have been reordered according to their centroids values. We are thus provided, for each pixel, with a sequence of discrete values as follows:

$$(NIR_1, R_6, G_3, NDVI_{16}) \rightarrow \cdots \rightarrow (NIR_{12}, R_3, G_{14}, NDVI_{19}) \qquad (1)$$

where $(NIR_1, R_6, G_3, NDVI_{16})$ means that the value of that pixel in the first image is in the first slice of near infra-red, in the 6^{th} slice of red, in the third slice of green and in the 19^{th} slice of $NDVI$.

5 Extracting Sequential Patterns from SITS

The preprocessing steps described in Sect. 4 provide us with a series of images where each pixel is described on a tuple of values. Let us consider the series of 3 images merely reduced to 4 pixels ($p1$ to $p4$) illustrated by Fig. 2. Each pixel in this figure is described on 3 values (corresponding to bands $B1$ to $B3$). With a minimum support of 100 %, there is no frequent pattern in these images (no "behaviour" corresponding to the whole set of pixels). With a minimum support of 50 %, however, we find two frequent behaviours:

1. $< (B1, white; B2, white)\ (B1, grey; B2, red) >$. This behaviour matches the sequences of values of pixels $p2$ on images 1 and 2 (or 3) and $p3$ on images 1 (or 2) and 3.
2. $< (B1, white; B2, white)\ (B1, white; B2, white) >$ (corresponding to $p1$ and $p3$ on images 1 and 2).

Let us note that, in the illustration above, patterns may be frequent even despite a lag in the images that support them.

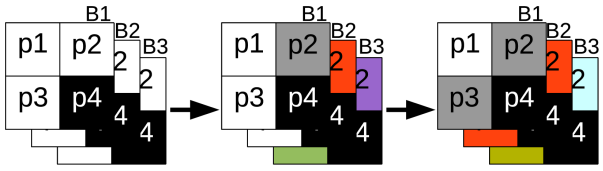


Fig. 2. A series of 3 images, with 4 pixels described on 3 bands.

Our goal is to extract sequential patterns, as described above. However, given the characteristics of our data, we find a large number of items (pixel values for one band) with a high support (say, more than 80 %). This has important

consequences on the discovery process. First, this will lead to numerous patterns which contain several occurrences of only one frequent item. Such patterns reveal non-evolutions such as $< (B1, white; B2, white)\ (B1, white; B2, white) >$ in our previous illustration and are not really informative. In our images, patterns with high support always correspond to geographic areas that did not change (these areas are majority).

To solve that issue, a naive approach would consist in lowering the minimum support in order to obtain patterns that correspond to changes (since the areas of changes are minority). Actually, specialists on this topic are interested in patterns that correspond to changes. Therefore, we need to extract patterns having lower support (say, between 1 % and 10 %).

However, let us consider $v_i b_j$ the i^{th} value on the j^{th} band. If the support of $v_i b_j$ is larger than 80 % then it is larger than any support below 80 %. Therefore, our extraction process will have to handle every frequent value during the discovery of frequent patterns. Unfortunately, these frequent values will flood the process with an intractable number of candidate and frequent patterns with two important consequences. First, the results are difficult, or even impossible, to obtain (because of the combination curse associated with frequent pattern extraction). Second, if the results can be obtained, they will be difficult to read and very few will be relevant because they will contain a lot of non-evolution patterns.

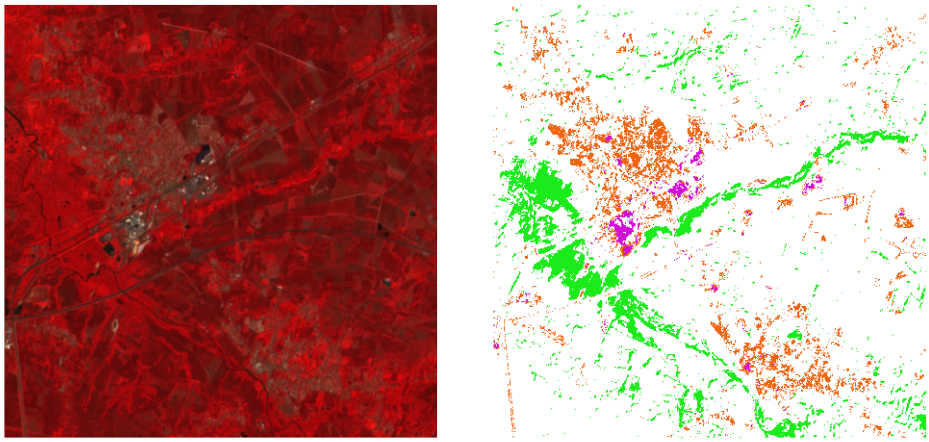
Therefore, we propose a frequent pattern extraction algorithm that is based on [2] with two important adjustments for SITS:

1. During the first step (discovery of frequent items) we only keep the items having threshold between a minimum and a maximum value. To that end, we have added a new support value to the process, which corresponds to a maximum support. Any item having support below the minimum or above the maximum value will be discarded.
2. During the remaining steps, we discard candidates that contain two successive identical values for a band. For instance, the candidate $< (B1, white)\ (B2, white) >$ is authorized, but not the candidate $< (B1, white)\ (B1, white) >$.

6 Experiments

Our images have a definition of 202,500 pixels (450x450). Once preprocessed, (as described in Sect. 4) each pixel takes values on 4 bands and our data contain a total of 28 millions values in the series. By applying the method described in Sect. 5 to the SITS illustrated in Fig. 1, we obtained patterns corresponding to thresholds between 5 % and 50 %. In this section, we report and describe three significant patterns selected from this result.

As we have illustrated in Fig. 2 a frequent pattern is extracted if it corresponds to the behaviour of a given number of pixels. When the pattern is found, we can retrieve the pixels whose series of values contain the pattern. These pixels may than be visualised (highlighted) as illustrated by Fig. 3(b). In this



(a) June 3 2006: image selected from the SITS

(b) Illustration of three selected patterns (one color per pattern)

Fig. 3. A sample of our results.

figure, each colour corresponds to a pattern selected from our SITS. Here is the geographic explanation of these patterns:

1. Pattern $< (IR, 1)\ (NDVI, 20) >$ is represented by the green dots in Fig. 3(b). It corresponds to swamps (wetlands) in the SITS. During winter, swamps are almost covered with water, resulting in a low infra-red level (slice 1) since water does not reflect light a lot. During summer, these swamps are not covered with water any more and light is reflected by the vegetation. Due to its high chlorophyll concentration (due to high irrigation), vegetation in summer has a very high level in NDVI (slice 20).
2. The orange dots represent pattern $< (R, 17)\ (R, 18;\ NDVI, 3) >$. It corresponds to urban areas that get denser (the number of residences has grew). Actually, urban areas (residences) have a high response in the red band. The level at the beginning of the pattern (slice 17) is highly likely to be the sign of a urban area. The following level (slice 18) shows a urban densification (slices 17 and 18 are separated by a radiometric increase of nearly 25 %), confirmed by a low level of NDVI (corresponding to almost no vegetation).
3. Pattern $< (NDVI, 2)\ (G, 20)\ (NDVI, 1) >$ is represented by the purple dots. This pattern corresponds to a densification of industrial areas (*e.g.* increase in the number of warehouses). In fact, industrial areas have high response in the green band and show very low values of NDVI. Furthermore, the decrease of NDVI (nearly 30 % from slice 2 to slice 1) shows that vegetation almost disappeared from these areas. Eventually, the maximum level of green is typical of flat roofs (*e.g.* corrugated iron) of industrial areas.

7 Conclusion

Our pattern extraction principle allowed us to find a significant number of relevant patterns such as the sample described in Sect. 6. Our patterns all have a geographic meaning. They correspond either to cyclic behaviours (swamps) or to long term evolutions through the dataset (densifications). Our technique is designed towards this specific extraction with a data mining process that takes into account a maximum support in the extraction. Indeed, when the support of a value is too high, it might lead to non-evolution patterns and numerous combinations. Thanks to our principle, this drawback is avoided and the discovered patterns are easy to read and understand.

8 Acknowledgement

The original source of publication is :
C. Fyfe et al. (Eds.): IDEAL 2010, LNCS 6283, pp. 45–52, 2010.
The original publication is available at www.springerlink.com.

References

1. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proceedings of the 11th International Conference on Data Engineering (ICDE'95). (1995) 3–14
2. Masseglia, F., Cathala, F., Poncelet, P.: The PSP Approach for Mining Sequential Patterns. In: Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery. (1998)
3. Bruzzone, L., Prieto, D.: Automatic analysis of the difference image for unsupervised change detection. IEEE Transactions on Geoscience and Remote Sensing **38**(3) (May 2000) 1171–1182
4. Todd, W.: Urban and regional land use change detected by using Landsat data. Journal of Research by the US Geological Survey **5** (1977) 527–534
5. Johnson, R., Kasischke, E.: Change vector analysis: a technique for the multispectral monitoring of land cover and condition. International Journal of Remote Sensing **19**(16) (1998) 411–426
6. Foody, G.: Monitoring the magnitude of land-cover change around the southern limits of the Sahara. Photogrammetric Engineering and Remote Sensing **67**(7) (2001) 841–848
7. Nielsen, A., Conradsen, K., Simpson, J.: Multivariate Alteration Detection (MAD) and MAF Postprocessing in Multispectral, Bitemporal Image Data: New Approaches to Change Detection Studies. Remote Sensing of Environment **64**(1) (1998) 1–19
8. Jha, C., Unni, N.: Digital change detection of forest conversion of a dry tropical Indian forest region. International Journal of Remote Sensing **15**(13) (1994) 2543–2552
9. Andres, L., Salas, W., Skole, D.: Fourier analysis of multi-temporal AVHRR data applied to a land cover classification. International Journal of Remote Sensing **15**(5) (1994) 1115–1121

10. Kennedy, R.E., Cohen, W.B., Schroeder, T.A.: Trajectory-based change detection for automated characterization of forest disturbance dynamics. Remote Sensing of Environment **110**(3) (2007) 370–386
11. Julea, A., Méger, N., Trouvé, E., Bolon, P.: On extracting evolutions from satellite image time series. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Volume 5. (2008) 228–231
12. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. (1967) 281–297