CLEARMiner: Mining of Multitemporal Remote Sensing Images

Aneesh Chandran
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Swathy Ramadas
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Abstract: A new unsupervised algorithm, called CLimate and rEmote sensing Association patterns Miner, for mining association patterns on heterogeneous time series from climate and remote sensing data, integrated in a remote sensing information system is developed to improve the monitoring of sugar cane fields. The system, called RemoteAgri, consists of a large database of climate data and low-resolution remote sensing images, an image pre-processing module, a time series extraction module, and time series mining methods. The time series mining method transforms series to symbolic representation in order to identify patterns in a multitemporal satellite images and associate them with patterns in other series within a temporal sliding window. The validation process was achieved with agro climatic data and NOAA-AVHRR images of sugar cane fields. Rules generated by the new algorithm show the association patterns in different periods of time in each time series, pointing to a time delay between the occurrences of patterns in the series analyzed, corroborating what specialists usually forecast without having the burden of dealing with many data charts. This new method can be used by agro meteorologists to mine and discover knowledge from their long time series of past and forecasting data, being a valuable tool to support their decision-making process.

Keywords: NOAA-AVHRR Images, Association Rules, Maximum Cross Correlation, Time Series Mining, Sequential Patterns

1. INTRODUCTION

The knowledge discovery, information mining methods and advances in computer technology have contributed to increase the access and application of remote sensing imagery. New technologies developed to be applied in the remote sensing area have increased its use in real applications. However, several users still have problems to deal with satellite images due to different and more sophisticated demands being imposed to them, as well as the fast growing in quantity and complexity of remote sensing data [1]. The knowledge discovery approach has been considered a promising alternative to explore and find relevant information on this huge volume of data. Some initiatives involving information and image mining have been accomplished through different techniques with reasonable results [2]–[4].

Association rules were proposed by Agrawall *et al.* [5] to solve the problem of discovering which items are bought together in a transaction. The number of rules discovered can be so large that analyzing the entire set and finding the most interesting ones can be a difficult task for the user. Then, Klemettinem *et al.* [9] proposed a method based on rule templates to identify interesting rules.

Instead of extracting features from images, other approaches work on computing measurements (indexes) from images generated by a combination of remote sensor channels that can be used to identify the green biomass, and soil temperature, for example. Thus, these indexes (measurements) can be extracted considering each pixel of multitemporal image data sets generating different time series. Time series are generated and studied in several areas, and data mining techniques have been developed to analyze them [5]–[7].

Mannila *et al.* [10] proposed a method to episodal sequential data mining that uses all frequent episodes within one sequence. Zaki [11] proposed the use of temporal constraints in transactional sequences. Harms *et al.* [12] defined methods that combine constraints and closure principles with a sliding window approach. Their objective was to find frequent closed episodes in multiple event sequence. In general, several techniques have been proposed to discover sequential patterns in temporal data in the last decade.

1.1 NOAA-AVHRR

The AVHRR is a radiation-detection imager that can be used for remotely determining cloud cover and the surface temperature. Note that the term surface can mean the surface of the Earth, the upper surfaces of clouds, or the surface of a body of water. This scanning radiometer uses 6 detectors that collect different bands of radiation wavelengths. The first AVHRR was a 4-channel radiometer, first carried on TIROS-N (launched October 1978). This was subsequently improved to a 5-channel instrument (AVHRR/2) that was initially carried on NOAA-7 (launched June 1981). The latest instrument version is AVHRR/3, with 6 channels, first carried on NOAA-15 launched in May 1998. The AVHRR/3 instrument weighs approximately 72 pounds, measures 11.5 inches X 14.4 inches X 31.4 inches, and consumes 28.5 watts power.
Measuring the same view, the array of diverse wavelengths, after processing, permits multi spectral analysis for more precisely defining hydrologic, oceanographic, and meteorological parameters. Comparison of data from two channels is often used to observe features or measure various environmental parameters. The three channels operating entirely within the infrared band are used to detect the heat radiation from and hence, the temperature of land, water, sea surfaces, and the clouds above them.

2. CLimate and rEmote sensing Association patteRns Miner (CLEARMiner)

A new unsupervised algorithm for mining association patterns on heterogeneous time series integrated to a remote sensing information system. The time series mining module was developed to generate rules considering a time lag. To do so, we define the constraint of time window to find association patterns that are extracted in two steps. First, the algorithm transforms multiple time series in a representation of patterns (peaks, mountains, and plateaus), with discrete intervals that maintain the time occurrence and represent phenomena on climate or remote sensing time series. In a second step, the algorithm generates rules that associate patterns in multiple time series with qualitative information. This algorithm-CLimate and rEmote sensing Association patteRns

Miner (CLEARMiner)-uses a sliding window value to find the rules that correspond to the number of patterns by window.

The algorithm quality is assessed using time series of agro meteorological data and multitemporal images from an important region of sugarcane production fields in Brazil. Sugarcane crops have expanded due to different reasons, such as, biofuel production, potential benefits to the environment as a possible way of mitigation of greenhouse gases emission, economic impact, among others. Although traditional ways to assess the sugar cane expansion exist, remote sensing images have been widely adopted to evaluate the direct land conversion to sugar cane. As sugarcane crops are cultivated on large fields, researchers have used satellites of medium and low spatial resolution, such as NOAA-AVHRR, 1 to identify areas for sugarcane expansion. We have also applied CLEARMiner to El Nino time series in order to discover their influence over precipitation distribution regime in regions of South America. In fact, both case studies are suitable to test the CLEARMiner algorithm since both experiments presuppose a relationship between series considering a time lag.

This algorithm works on multiple time series of continuous data, identifying patterns according to a given relevance factor (r) and a plateau length (l) thresholds. In its last step, the algorithm associates patterns according to a temporal sliding window that corresponds to the number of patterns. The number of patterns decreases when the tuning parameters increase, as the experiments showed. Patterns can be seen as discrete intervals that allow the association between series. CLEARMiner presents rules in two formats: short and extended. Short rules are easier to understand, but they are not sufficient to visualize the peak amplitudes and the length of the plateaus. Therefore, the algorithm also presents rules in extended format including details of the values variation and time intervals.

3. ARCHITECTURE OF REMOTE AGRI

Before applying data mining techniques in remote sensing imagery, it is necessary to submit images to the preprocessing process. The knowledge discovery process in information mining systems involves three main phases: data preparation, data mining, and presentation of knowledge. Geometric correction combines indirect navigation and spacecraft attitude error estimation. After that, the maximum cross correlation technique can be used to detect the geographic displacement between the base image and the target one.

Module 1 corresponds to the image georeferencing step executed in batch mode by NAVPro; Module 2 is executed by SatImagExplorer which was proposed to extract values or compute indexes from multitemporal images generating time series for each pixel of the image; and Module 3 refers to time series mining module (CLEARMiner) developed to associate climate data with indexes extracted from NOAA-AVHRR images.

3.1 System Prototype

The system prototype consists of three major components as shown in Fig -1:

- image georeferencing module
- time series extraction module
- time series mining method

The first module to be executed in the RemoteAgri system corresponds to the image georeferencing process, as is presented in figure 2.1. This module is composed of several Cshell scripts that call the subroutines of NAV system in batch mode to accomplish necessary tasks, such as:

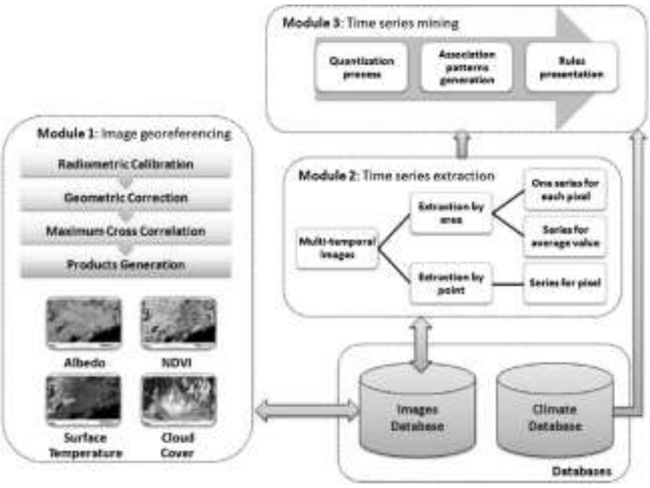


Fig -1: Schematic diagram of the multitemporal images mining system—RemoteAgri

- conversion from raw format to an intermediary;
- radiometric calibration;
- geometric correction;
- identification of pixels classified as cloud.

The georeferencing module allows users to generate four different synthesis images: albedo, NDVI, surface temperature, and cloud cover for a specific region as shown in Fig. 1.

As the volume of images is huge, an extraction module called SatImagExplorer was proposed to perform it faster and in a more flexibly way. The second module extracts values or computes the index from the images opened. Then, it generates a time series computing the index values for all images using the same coordinate (latitude/longitude) of the region. In addition to the direct interaction with the system interface, users can also extract time series using a vector of coordinates that defines the desired region. Time series extracted from multitemporal images SatImagExplorer are then mined in order to discover patterns or association patterns. The last module refers to time series mining developed to associate climate data to indexes extracted from NOAAAVHRR images.

3.2 Maximum Cross Correlation Method

The maximum Cross Correlation (MCC) method is used to automatically compute the satellite attitude parameters required to geometrically correct images to this base image. The MCC method detects the geographic displacements between the base image and the target image[5]. These image displacements are then used to compute the roll, pitch, and yaw attitude parameters. This approach requires the base image to have minimal cloud cover to maximize the potential sites for the computation of image offsets. In the actual application of the base image to the navigation of subsequent images, several levels of cloud detection are applied to ensure that clouds do not influence the image correction calculations.

The base image must be registered to the exact same grid as the target images and must be as cloud free as possible. A second requirement is that the radiance distributions of the base and target images be similar. The widely varying illumination conditions that exist for different orbits prevent the use of the reflected channels for this algorithm.

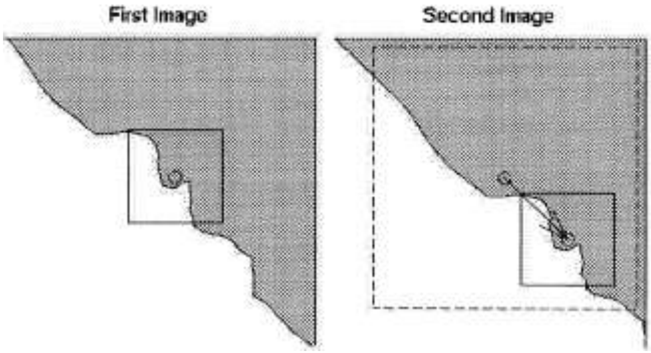


Fig -2: MCC Method applied to two sequential satellite images

4. QUANTZATION PROCESS

The process of time series mining is divided into three parts as shown in figure 2.1. The Quantization process module receives as input a set of remote sensing and climate time series. The time series is defined as a sequence of pairs (ai, ti) with i = 1. . . n, i.e., S = [(a1, t1), . . . , (ai, ti), . . . , (an, tn)] and (t1 < . . . < ti < . . . <tn), where each ai is a data value, and each ti is a time value in which ai occurs. Each pair (ai, ti) is called an evente. A set of events E contains n events of type ei = (ai, ti) for i = 1. . . n. Each ai is a continuous value. Each ti is a unit of time that can be given in days, months or years. Given two sequences S1 and S2, the values ti of both sequences must be measured in the same time unit.

A set of consecutive ei, i.e., Se = (ei, ei+1. . . ek), where ei = (ai, ti) for ti >= t1 and tk <= tn is called the event sequence Se. The number of elements ei in the event sequence depends on the difference between events given by di = (ai+1 - ai) (1st step in figure 3.2), and a given d parameter whose default value is set by the algorithm.

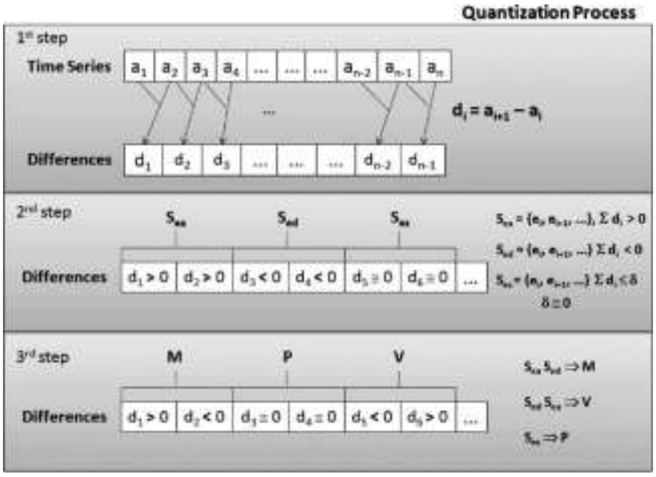
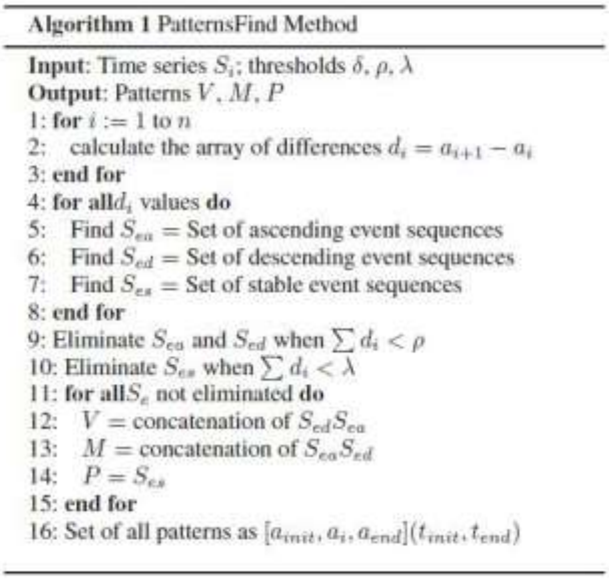

Quantization Process
1st step
Time Series
a1 a2 a3 a4 ... ... an-2 an-1 an
...
di = ai+1 − ai
Differences
d1 d2 d3 ... ... dn-2 dn-1
2nd step
Sea
Sed
Ses
Sea = {ei, ei+1, ...}, Σ di > 0
Sed = {ei, ei+1, ...} Σ di < 0
Ses = {ei, ei+1, ...} Σ di ≤ δ
δ ≅ 0
Differences
d1 > 0 d2 > 0 d3 < 0 d4 < 0 d5 ≅ 0 d6 ≅ 0 ...
3rd step
M
P
V
Sea Sed ⇒ M
Sed Sea ⇒ V
Ses ⇒ P
Differences
d1 > 0 d2 < 0 d3 ≅ 0 d4 ≅ 0 d5 < 0 d6 > 0 ...

Fig -3: Representation of the three steps of the quantization process executed by time series mining module (CLEARMiner).

V (valley) corresponds to a pattern defined as the concatenation of a descending event sequence and an ascending event sequence (i.e., V = $S_{ed}S_{ea}$). P (plateau) represents a kind of pattern described as a stable event sequence (i.e., P = S_{es}), while M (mountain) indicates a pattern generated by the concatenation of an ascending event sequence and a descending event sequence (i.e., M= $S_{ea}S_{ed}$). Figure 3.3 presents an example of a pattern V. In real data, V can be observed when a sharp drop in the minimum temperature occurs. Algorithm 1 presents the main idea used to convert a time series in a pattern sequence of V, P, and M.



Algorithm 1 PatternsFind Method

```
Input: Time series $S_i$; thresholds $\delta, \rho, \lambda$
Output: Patterns $V, M, P$
for $i := 1$ to $n$
    calculate the array of differences $d_i = a_{i+1} - a_i$
end for
for all$d_i$ values do
    Find $S_{ea}$ = Set of ascending event sequences
    Find $S_{ed}$ = Set of descending event sequences
    Find $S_{es}$ = Set of stable event sequences
end for
Eliminate $S_{ea}$ and $S_{ed}$ when $\sum d_i < \rho$
Eliminate $S_{es}$ when $\sum d_i < \lambda$
for all$S_e$ not eliminated do
    $V$ = concatenation of $S_{ed}S_{ea}$
    $M$ = concatenation of $S_{ea}S_{ed}$
    $P = S_{es}$
end for
Set of all patterns as $[a_{init}, a_i, a_{end}](t_{init}, t_{end})$
```

The algorithm concatenates consecutive sequences Sea and Sed to generate an M pattern, Sed and Sea to generate a V pattern, and Ses to generate P patterns.

5. ASSOCIATION PATTERNS GENERATION

A pattern in one time series can be associated to patterns in other time series. Consider an association pattern as an expression of the form Si[a] => Sj [b], where Si and Sj are different time series (for example, rainfall series), a and b are frequent patterns, such as M, V , or P.

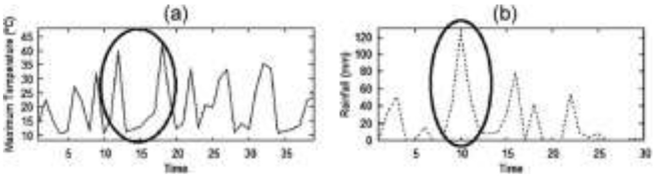


Fig -4: (a) Pattern of type V is similar to negative peaks. (b)Pattern of type M is equivalent to a positive peak.

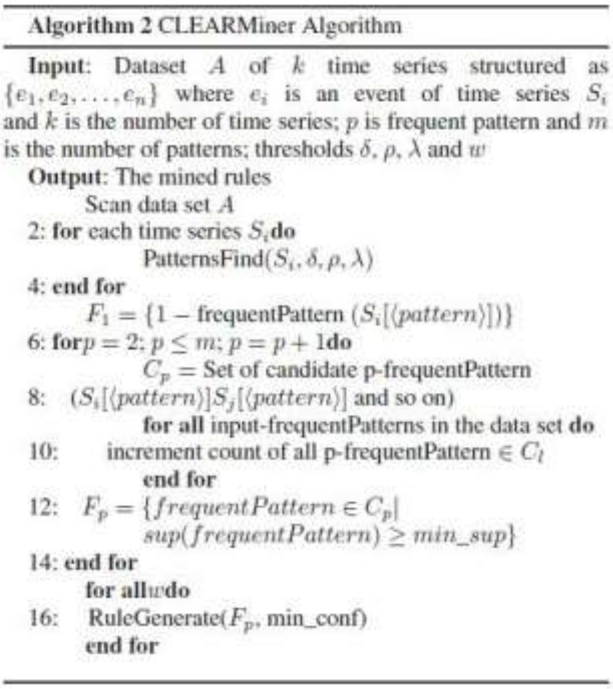
The support of Si[a] => Sj [b] represents the frequency of occurrences and is given by

$$support = \frac{fr\,(S_i[\alpha], S_j[\beta])}{T}$$

The confidence measure indicates the percentage of all patterns in Si and Sj containing Si[a] that also contain Sj [b]. The confidence for the rule Si[a]) Sj [b] is given by,

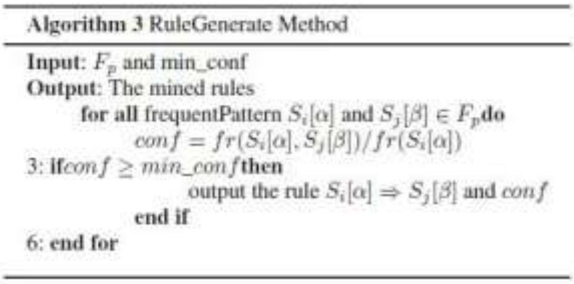
$$conf = \frac{fr\,(S_i[\alpha], S_j[\beta])}{fr\,(S_i[\alpha])}$$

Algorithm 2 shows the pseudocode for CLEARMiner. The CLEARMiner algorithm calculates j-frequentPatterns for each time series. The algorithm only stores j-frequentPatterns greater than the *min_sup* threshold defined by the user.



```
Algorithm 2 CLEARMiner Algorithm
Input: Dataset $A$ of $k$ time series structured as $\{e_1, e_2, \dots, e_n\}$ where $e_i$ is an event of time series $S_i$ and $k$ is the number of time series; $p$ is frequent pattern and $m$ is the number of patterns; thresholds $\delta$, $\rho$, $\lambda$ and $w$
Output: The mined rules
    Scan data set $A$
  for each time series $S_i$ do
        PatternsFind($S_i, \delta, \rho, \lambda$)
  end for
    $F_1 = \{1 - \text{frequentPattern } (S_i[\langle pattern \rangle])\}$
  for $p = 2; p \leq m; p = p + 1$ do
        $C_p$ = Set of candidate p-frequentPattern
    ($S_i[\langle pattern \rangle] S_j[\langle pattern \rangle]$ and so on)
        for all input-frequentPatterns in the data set do
      increment count of all p-frequentPattern $\in C_l$
        end for
    $F_p = \{frequentPattern \in C_p |$
        $sup(frequentPattern) \geq min\_sup\}$
  end for
    for all $w$ do
    RuleGenerate($F_p$, min_conf)
    end for
```

For each frequent pattern in F, the algorithm calculates, via the RuleGenerate Method, the confidence value (line 2- Algorithm 3). If confidence is greater than minconf, it generates rules (lines 3 to 5-Algorithm 3).



```
Algorithm 3 RuleGenerate Method
Input: $F_p$ and min_conf
Output: The mined rules
      for all frequentPattern $S_i[\alpha]$ and $S_j[\beta] \in F_p$ do
          $conf = fr(S_i[\alpha], S_j[\beta]) / fr(S_i[\alpha])$
3: if $conf \geq min\_conf$ then
              output the rule $S_i[\alpha] \Rightarrow S_j[\beta]$ and $conf$
          end if
6: end for
```

6. RULES PRESENTATION

The algorithm presents the rules in two formats to better visualize them: short (the succinct way) and extended (those with more details and time stamp)[10]. The short format is more succinct and easier to be analyzed. However, it contains no information about the context in which the phenomenon occurred.

This rule indicates that the pattern [ai, ak, an] occurred in the period (tinit1 - tend1) for the time series S1, which is associated to the pattern [aj, al, am] occurred in the period (tinit2 - tend2) for the series S2 with tinit1 tinit2 and tend1 tend2. Thus, the user can analyze rules in the short format to verify correlations between time series and to use the extended format to obtain more details. An example with real data is presented in Fig 3.

7. ADVANTAGES

The results are shown by comparing the CLEARMiner algorithm with two classical and baseline algorithms, apriori [7] and the generalized sequential pattern (GSP) algorithm.
Both algorithms were performed in the Weka platform. As the two algorithms work with discrete data, we compared only the rules generation. The data sets used to run apriori and GSP were quantized by CLEARMiner to avoid distortions that could be caused by different quantization processes. The apriori algorithm mined few rules and did not consider time of occurrences.

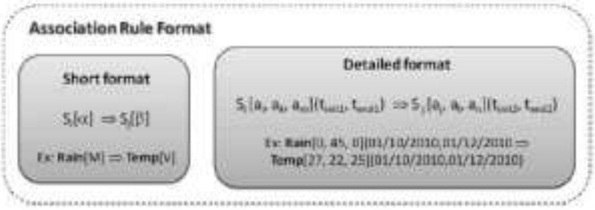

Association Rule Format
Short format
$S_i[\alpha] \Rightarrow S_j[\beta]$
Ex: Rain[M] ⇒ Temp[V]
Detailed format
$S_i[a_i, a_k, a_n](t_{init1}, t_{end1}) \Rightarrow S_j[a_j, a_l, a_m](t_{init2}, t_{end2})$
Ex: Rain[0, 45, 0](01/10/2010,01/12/2010) ⇒
Temp[27, 22, 25](01/10/2010,01/12/2010)

Fig -5: Examples of rules in short and extended format in time series mining module.

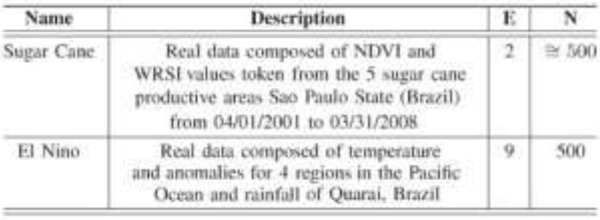
The GSP algorithm scans the database several times to generate a set of candidate k-sequences and to calculate their support. We executed the GSP algorithm with min_sup =0.2. For minsup values above 0.2, the GSP algorithm in Weka did not work properly. The sequences mined by GSP are similar to the rules generated by CLEARMiner. However, both algorithms (Apriori and GSP) do not keep information about the time occurrence of the events.
CLEARMiner generates rules in an extended format, which can be used to obtain more details about the correlation between time series.

Another advantage of this method is the quantization process that is executed as a first step. This quantization generates a representation that encompasses the semantics meaningful for climate and agroclimate time series. The criteria to quantize time series are based on phenomena that are observed by meteorologists and agrometeorologists and impact the environment.

8. APPLICATIONS

The multitemporal NDVI images from NOAA-AVHRR were studied, covering the scene with orbit/point 220/75 of Landsat satellite. We have selected regions located in Sao Paulo, which is responsible for the majority of sugar cane production in the country. Sugar cane crops are cultivated in plain relief. The climate of this region presents fluctuations in temperature during the rainy season: October to March. The results of experiments were performed on two real data sets to evaluate and validate the proposed algorithm. The results from such experiments followed the specialists' expectations and helped on tuning the algorithms' parameters. Table I presents a summary of the data sets used, giving their dimensions number (E) and the size of time series (N).

Table -1: Definition of datasets that was used to evaluate the performance of CLEARMiner



Name	Description	E	N
Sugar Cane	Real data composed of NDVI and WRSI values token from the 5 sugar cane productive areas Sao Paulo State (Brazil) from 04/01/2001 to 03/31/2008	2	≅ 500
El Nino	Real data composed of temperature and anomalies for 4 regions in the Pacific Ocean and rainfall of Quarai, Brazil	9	500

Two main applications are:

- Mining NDVI and WRSI Time Series From Sugar Cane Regions
- Mining Time Series of Rainfall and Anomalies Related to El Nino

9. CONCLUSION

The system, called RemoteAgri, consists of a large database of climate data and low-resolution remote sensing images, an image preprocessing module, a time series extraction module, and time series mining methods [1]. The preprocessing module was projected to perform accurate geometric correction, what is a requirement particularly for land and agriculture applications of satellite images. The time series extraction is accomplished through a graphical interface that allows easy interaction and high flexibility to users. The time series mining method transforms series to symbolic representation in order to identify patterns in a multitemporal satellite images and associate them with patterns in other series within a temporal sliding window.

The results show that the algorithm detects some association patterns that are known by experts, as expected, indicating the correctness and feasibility of the proposed method. Moreover, other patterns detected using the highest relevance factors are coincident with extreme phenomena as many days without rain or heavy rain as the specialists suppose to. The mined rules for the relevance patterns indicate a relation between series, allowing these patterns (phenomena) happen in different intervals of time. This method can be used by agrometeorologists to mine and discover knowledge from their long time series of past and forecasting data, being a valuable tool to support their decision-making process.

10. REFERENCES

[1] Luciana Alvim S. Romani, Ana Maria H. de Avila "A New Time Series Mining Approach Applied to Multitemporal Remote Sensing Imagery", Communications of the ACM, 21:140-150, 2013.

[2] R.M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli. "Information mining in remote sensing image archives: System concepts", 2003.

[3] J. Li and R. M. Narayanan "Integrated spectral and spatial information mining in remote sensing imagery", Communications of the ACM, 54(8):62–71, 2011.

[4] Witold Pedrycz G. B. Yingxu Wang. "Information mining in remote sensing image archives: System evaluation", IEEE Trans. Geosci. Remote Sens, page 188-199, 2005.

[5] W. Emery, D. G. Baldwin, and D.Matthews, "Maximum cross correlation automatic satellite image navigation and attitude corrections for open ocean image navigation", IEEE Proceedings.

[6] J. C. D. M. Esquerdo, J. F. G. Antunes, D. G. Baldwin., "An automatic system for AVHRR land surface product generation," 2003.

[7] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in Proc. 4th Int. CFDOA, Chicago, IL, 1993, pp. 69-84.

[8] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule discovery from time series," in Proc. 4th ICKDDM, New York, 1998, pp. 16-22.

[9] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in Proc. ICEDT, Avignon, France, 1996, pp. 3-17.

[10] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkano, "Finding interesting rules from large sets of discovered association rules," in Proc. CIKM, Gaitherburg, MD, 1994, pp. 401-407.

[11] H. Mannila, H. Toivonem, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," Data Mining Knowl. Discovery, vol. 1, no. 3, pp. 259-289, 1997.

[12] M. Zaki, "Sequence mining in categorical domains: Incorporating constraints," in Proc. CIKM, Washington, DC, 2000 pp. 422-429.

[13] S. K. Harms, J. Deogun, J. Saquer, and T. Tadesse, "Discovering representative episodal association rules from event sequences using frequent closed episode sets and event constraints," in Proc. ICDM, San Jose, CA, 2001,pp. 603-606.

[14] J. Wang and J. Han, "BIDE: Efficient mining of frequent closed sequences," in Proc. ICDE, 2004, pp. 79-90