Time Series Numerical Association Rule Mining for assisting Smart Agriculture

Iztok Fister Jr. University of Maribor Maribor, Slovenia iztok.fister1@um.si Iztok Fister University of Maribor Maribor, Slovenia iztok.fister@um.si Sancho Salcedo-Sanz Universidad de Alcalá Alcalá de Henares, Spain sancho.salcedo@uah.es

Abstract-Smart agriculture is a modern paradigm that incorporates modern technologies ranging from cloud computing, big data, Internet of Things, or data mining, among others. The final objective of smart agriculture is to assist the classical farming process in automating some essential steps, and making agriculture more sustainable. In agriculture, different time-series data allow the application of apply machine learning methods in order to solve different problems. Currently, most of the techniques were developed for classification or prediction tasks. On the contrary, this paper sheds light on the potential of numerical association rule mining for time series data. In line with this, a novel method for mining time series numerical association rules is proposed on time-series data collected by monitoring parameters detected during a growth of plants. Experiments reveal that time series numerical association rule mining could be a suitable method for mining time-series data in agriculture and related areas.

Index Terms—association rule mining, numerical association rule mining, optimization, smart agriculture

I. INTRODUCTION

Global warming, floods, drastic climate changes, heatwaves, and soil contamination are just a few examples of environmental challenges [17], [20], [26], that have changed agriculture in the last decades. They have launched the question about how to effectively produce a sustainable amount of food [3], [4]. Moreover, the food prices are highly rising on a global scale. Consequently, more food is producing on the one hand, while healthy food is nowadays the big goal of many farmers around the world, on the other.

Fortunately, farmers are nowadays not only interested in a conventional way of farming, but they are also searching for modern solutions allowing them: to automatize some key steps in their farming process, to have a big crop following to the principles of ecological agriculture, and to help them follow the guidelines of sustainable farming. The sustainable farming is also the strategic goal in the policies of different countries worldwide. In addition, consumer electronics and modern Information Technology (IT) are now becoming affordable almost to the everyone. Nowadays, buying a small computer in the credit-card size called Raspberry-Pi costs less than 20 Eur, while different sensors for agriculture, enabling it to act in the real environment, can even be cheaper. These sensors play

vital role in the agriculture technological revolution, because of enabling farmers to monitor actual parameters arisen during the plant growth in a cheap way. Obviously, the parameters can be digitalized, saved into databases and later analyzed using sophisticated methods.

Machine Learning (ML) methods [14], [19] present the one of the most significant analyzing tools, nowadays. These methods can be used for solving the classification, regression or prediction tasks, among others. Moreover, Association Rule Mining (ARM) is an advanced ML method [1] intended to discover dependencies between attributes in a transaction database. An extension to the ARM is Numerical Association Rule Mining (NARM), which mitigates several drawbacks of canonical ARM [10]. In NARM, we can deal with numerical and categorical data concurrently, while numerical data needs to be discretized in bare ARM. This can result in incorporating the noise in data and mined association rules that can have more considerable discrepancies and may not show an accurate picture.

Continuous monitoring of plant health or plant parameters is crucial in ecological agriculture [8], [13]. With the continuous monitoring of plant, we can obtain invaluable data that have been observed at different points in time. This leads us to a domain of time series analyzing that looks for the answers on mathematical and statistical questions posed by time correlations [22]. As a possible solution, the Time Series Numerical Association Rule Mining (TS-NARM) algorithm is proposed, able to extract the hidden knowledge in time series data and potentially adapt the agriculture process by the new insights. Thus, we are interested in model the relation between features, like the ambient temperature, soil moisture and plant, in order to obtain association rules that reflect the main relations between them.

To the knowledge of authors, TS-NARM has not been applied regularly to the ML applications in smart agriculture. There exist several use cases of bare ARM, but we were unable to find any specific TS-NARM use cases in this domain. This paper tries to contribute to smart agriculture with the innovate use of NARM to extract hidden knowledge from different features that were monitored by growing process of a single tree. We show that TS-NARM algorithms can find some hidden relationships between attributes and thus contribute to better understanding of different parameters involved in the plant growth. With this knowledge the process can be improved or adapted to the new level. More detailed contributions of this paper are as follows:

- a proposal of experimental framework, i.e., hardware and software solution for monitoring a plant and capturing parameters,
- · a new numerical association rule for time series mining,
- an evaluation of captured parameters,
- an explanation of obtained results.

The structure of the remainder of the paper is as follows: In Section II, background information needed for understanding the paper are discussed. Section III reveals a development of the TS-NARM. The goal of Section IV is to apply the proof of concept in order to validate it. Section V concludes the paper by summarizing of the performed work and outlining the directions for the future.

II. BACKGROUND

This section is devoted to discuss the background information to complete the paper. In line with this, it is divided into three subsections: (1) basics of the ARM, (2) introductory concepts about TS-NARM, and (3) review of the existing ARM use-cases in smart agriculture.

A. Association Rule Mining

The ARM problem can be defined as follows: Let us suppose a set of objects $O = \{o_1, \ldots, o_n\}$ and a transaction set D, where each transaction T is a subset of objects, i.e. $T \subseteq O$. Then, an association rule can be defined as an implication:

$$X \Rightarrow Y,\tag{1}$$

where $X \subset O$, $Y \subset O$, in $X \cap Y = \emptyset$. The following two measures are defined for evaluating the quality of the association rule [1]:

$$conf(X \Rightarrow Y) = \frac{n(X \cup Y)}{n(X)},$$
 (2)

$$supp(X \Rightarrow Y) = \frac{n(X \cup Y)}{N},$$
 (3)

where $conf(X \Rightarrow Y) \geq C_{min}$ denotes confidence and $supp(X \Rightarrow Y) \geq S_{min}$ support for the association rule $X \Rightarrow Y$. Thus, N in Equation (3) represents the number of transactions database D, and n is the number of repetitions of a particular rule $X \Rightarrow Y$ within D. Here, C_{min} denotes the minimum confidence and S_{min} the minimum support. This means that only those association rules with confidence and support higher than C_{min} and S_{min} are taken into consideration, respectively.

B. Time Series Association Rule Mining

Time Series Association Rule Mining (TS-ARM) is a new paradigm, which treats a transaction database as a time series data. In line with this, the formal definition of the ARM problem needs to be slightly redefined. In our study, the transaction database is defined as a set $D = \{d_1, \ldots, d_n\}$, where elements d_i for $i = 1, \ldots, n$ represent a sequence of transaction observed in some time interval $t = (t_0, t_1)$. Thus, each transaction consists of a set of objects $O = \{o_1, \ldots, o_m\}$. Then, the association rule is defined as an implication:

$$X(t) \Rightarrow Y(t),\tag{4}$$

where $X(t) \subset O$, $Y(t) \subset O$, and $X(t) \cap Y(t) = \emptyset$. The measures of support and confidence are redefined as follows:

$$conf_T(X(t)) \Rightarrow Y(t)) = \frac{n(X(t) \cup Y(t))}{n(X(t))},$$
 (5)

$$supp_T(X(t)) \Rightarrow Y(t)) = \frac{n(X(t) \cup Y(t))}{N(t)}, \qquad (6)$$

where $conf_T(X(t) \Rightarrow Y(t)) \ge C_{\max}$ and $supp_T(X(t) \Rightarrow Y(t)) \ge S_{\max}$ denotes the confidence and support of the association rule $X(t) \Rightarrow Y(t)$ within the time interval $t = (t_0, t_1)$.

Let us notice that a transaction database is divided into more slices using the time intervals. However, these slices can be treated either absolutely or relatively. In the first case, the time interval is uniquely selected on the timeline, while in the second, the timeline is divided into periods that repeat at every time interval per period simultaneously (e.g., day, week, year).

C. ARM use cases in smart agriculture

Data Mining is a vital part of discovering new knowledge/insights from data. Data mining can be found in almost all areas, and agriculture is not an exception. Probably one of the first demonstrations of using ARM in the domain of agriculture was presented in [5] where the authors showed the potential of the WEKA software tool on an agricultural dataset. On the other hand, the work [24] depicts the framework of the pest management system using data mining techniques, where association rule mining is a part of this framework. The framework provided historical data, current and recommended pest and pesticide information, and simulated pest models up to the farm level. Interestingly, the authors in [21] propose a conceptual framework to investigate how farmers' dynamic agricultural activities under different socio-economic conditions affect the water-energy-food (WEF) systems. A proposed model consisting of an Association Rule Mining (ARM) and Agent-Based Model (ABM) has been employed to inspect farmers' activities. In a slightly different way, authors in [18] used the association rule mining to analyze the knowledge about sweet potato (Ipomoea batatas L.) in Slovenia. Association rule mining of agricultural machinery maintenance was used for data analysis of the potential relation to prevent failure and promote working effect [12].

Finally, it is worth mentioning that several review papers also outlined the role of association rule mining in agriculture [16], [25].

III. DEVELOPMENT OF THE PROPOSED METHOD

The proposed method consists of the following four phases:

- data collection,
- data preprocessing,
- data processing using TS-ARM,
- data explanation.

In the remainder of the paper, these phases are discussed in detail.

A. Data collection

Data collection is devoted to aggregate data achieved from different agriculture sensors. In general, six different sensors are typically applied as follows: pH, GPS, temperature, asset, moisture, and accelerometer sensors. These sensors can be attached to an acquiring system either wired or wireless. The format of data, they transmit to the system in predefined time intervals, is simple, i.e., the specific value of measured feature as a floating-point values.

Mathematically, data collection phase can be described as follows: Let us assume *n*-sensors for agriculture are devoted for ambient condition monitoring. Data from these sensors are acquired on the acquiring system into a tuple:

$$TU = \langle sense_1, \dots, sense_n, t \rangle, \tag{7}$$

where $sense_i$ for i = 1, ..., n are the floating-point values denoting the ambient condition of *i*-th sensors and *t* represents a timestamp.

B. Data preprocessing

Data preprocessing is usually one of the most critical steps in the whole data science process. Data preprocessing can be defined as a set of methods that enhance the overall quality of the raw data and try to enrich it [9]. Well-defined and wellprepared data has a huge influence on the overall accuracy of results produced by data science pipelines. Essentially, two tasks are required in time series data preprocessing phase:

- time frame creation,
- feature extraction.

The first preprocessing task enables grouping the data in time frames, while the second is devoted to data enrichment.

1) Time frame creation: The main goal of this task is to group corresponding time series data in a specified time frame. In this phase, a specific number of tuples TU are aggregated into the so-called time frame TF, in other words:

$$TF = \langle SE_1, \dots, SE_n, t_N \rangle, \tag{8}$$

where each element of TF_i for i = 1, ..., n is expressed as:

$$SE_i = \frac{1}{N-1} \sum_{t=t_0}^{t_N} sense_i^{(t)},$$
 (9)

and t_0 and t_{N-1} denote a start and end of the observed time interval.

Indeed, the measured floating-point values obtained from specific sensor that measures the particular ambient condition are averaged during the predefined time intervals and aggregated into time frames. However, the aim of this frame construction is also to avoid noise in data transmitting from the sensors and thus, to increase the robustness of the proposed method.

2) Feature extraction: The time frames TF consist of raw parameters representing a base from which separate features can be defined. The features are building blocks from which each transaction database D consist. The transaction database enters into an ML pipeline.

Due to the lack of features, the NARM methods are not capable to produce any specific insights. In order to break this gap, we must enrich data from initial dataset using additional features that show a better outlook on collected data. Finally, a set of features (objects) O is defined with elements o_i for i = 1, ..., n representing features. Each transaction T in transaction database denotes the appropriate time frame obtained during the last phase.

C. TS-ARM algorithm

The TS-ARM algorithm is defined as an evolutionary algorithm [6], [7] and each individual is represented as a vector:

$$\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{t}\},\tag{10}$$

where the elements are represented as quadruples $\mathbf{x}_i = \langle \pi_i, Lb_i, Ub_i, Th_i \rangle$ for i = 1, ..., n, and is a pair $\mathbf{t} = \langle t_0, t_1 \rangle$. The meaning of quadruple elements is the following: π_i denotes a position in the permutation of features, Lb_i the lower and Ub_i the upper bound of the numeric interval, while the Th_i the threshold determining the control point of the corresponding association rule.

The fitness function is calculated according to the following equation:

$$f(\mathbf{x}_i^{(t)}) = \frac{\alpha \cdot supp(X \Rightarrow Y) + \beta \cdot conf(X \Rightarrow Y)}{\alpha + \beta}, \quad (11)$$

where α , and β , denote weights, $supp(X \Rightarrow Y)$ and $conf(X \Rightarrow Y)$ represent the support and confidence of the observed association rule. In practice, any variants of stochastic population-based nature-inspired algorithm could be used in role of TS-ARM algorithm.

IV. PROOF-OF-CONCEPT

The goal of our experimental work was to prove that the proposed method is the proper solution for using in agriculture practice. In line with this, all four phases of the proposed method are tested, while the obtained results are illustrated in details.

For the proof of concept, we set up our hardware in the soil of an Aloe Vera plant for ten days (Fig. 1). Fig. 1.a presents a Nectarine plant, while Fig 1.b presents two Aloe Vera plants in a pot. Nectarine plant is located outside of the building, while



(a) Nectarine plant (Outside)

Fig. 1. Test scenarios.

Aloe Vera plant is located inside during the Winter season in Europe¹

Note that sensors captured ambient condition data every 5 seconds and stored it in a data class model. Every 6 minutes, data were dumped into a JSON file. During the test scenario, the Aloe Vera plant was watered two times.

An elementary hardware environment consisted of the following hardware components, in our study (Fig. 2):

- Raspberry Pi Model B,
- Adafruit STEMMA soil sensor,
- STEMMA QT/Qwiic JST SH 4-pin Cable with Female Sockets, and
- solar power bank for monitoring the plant outside.

The Raspberry Pi Model B computer plays a role of acquiring system, to which a soil sensor is connected.



Fig. 2. Hardware solution

The Adafruit STEMMA soil sensor is capable to measure moisture as well as temperature and it is the primary sensor

¹Aloe Vera plants will also be moved outside in the beginning of May in central Europe.



(b) Aloe Vera plants inside

component for data collection. The STEMMA QT/Qwiic JST SH 4-pin Cable enables connecting the sensors to Raspberry Pi, while the solar power bank is powering our system.

The data collected by the acquiring system are represented just using three parameters:

- an ambient temperature,
- a moisture,
- a timestamp.

The program, which ran on the Raspberry PI device, was written in Python and captured measurements every 5 seconds. Data were stored in a JSON format, allowing faster data preprocessing. Experiments were run on an HP Z240 workstations computer with the following setup: Linux Mint 19.1 Tessa OS, 16 GB of physical memory, Intel Xeon E3-1245 processor.

In the data preprocessing phase, we first merged measurements in 6-minute frames, to reduce any potential noise in the data. After this step, a transaction database was created using the feature extraction as presented in Table I. The transaction database served as an input to the association rule mining process. ARM steps were implemented in Python using the NiaPy library [27].

The following five population-based nature-inspired algorithms were added in comparison in processing phase: DE [23], PSO [15], FA [28], jDE [2], GA [11]. Implementations of all the mentioned algorithms were taken from NiaPy library [27]. The default parameters as specified in NiaPy examples were used, while the number of function evaluations for all algorithms was set to 10000 function evaluations. All algorithms, in comparison, ran ten independent runs.

The results of experiments are collected in Table II, where average support and average confidence metrics are presented along with the average antecedents/consequents that appeared in the rules. Additionally, an average number of transactions selected as a part of the transactions database with the average number of association rules are also presented. As can we see from the table, the DE algorithm discovered the most rules among all algorithms. Rules discovered by the DE algorithm also had the highest confidence average. The average support of the DE algorithm was very close to the highest support

TABLE I: Extracted features and their description.

Feature	Attribute domain	Short description	
AVG-TEMPERATURE	NUMERIC	Average temperature of data in specific time frame	
MAX-TEMPERATURE	NUMERIC	Max temperature of data in specific time frame	
MIN-TEMPERATURE	NUMERIC	Min temperature of data in specific time frame	
AVG-MOISTURE	NUMERIC	Average moisture of data in specific time frame	
MAX-MOISTURE	NUMERIC	Max moisture of data in specific time frame	
MIN-MOISTURE	NUMERIC	Min moisture of data in specific time frame	
DIFF-TEMPERATURE	NUMERIC	Temperature difference between first and last time point in time frame	
DIFF-MOISTURE	NUMERIC	Moisture difference between first and last time point in time frame	

TABLE II: Results of experiments

Measure/Algorithm	Avg. support	Avg. Confidence	Avg. antecedent	Avg. consequent	Number of transactions	Avg. number of rules
DE	0.61	0.79	1.57	1.71	822	4837
jDE	0.5	0.69	1.39	1.78	932	2797
PSO	0.65	0.67	1.05	1.5	653	1965
GA	0.16	0.47	1.88	1.84	674	2128
FA	0.2	0.49	1.69	2.03	828	2080

average obtained by the PSO algorithm. Contrary to the DE algorithm, the PSO algorithm discovered *approx* 60% less of the rules. It may indicate that the DE algorithm is much more exploratory than PSO when solving the association rule mining problem. The other evolutionary methods tested obtain worse results.

A. Discussion

In our preliminary work, we did not pay any special attention to the accuracy and sensitivity of accompanying sensors we used for our experiments. Our primary goal was to build a very cheap solution which can capture "reliable" data. We observed some discrepancies in consecutive measurements when using these sensors. However, these discrepancies, which can also be denoted as anomalies, may easily be filtered out from the dataset in the future. On the other hand, the whole dataset produced solely on actual measurements proved very versatile for applying a novel association rule mining method that was also a part of this study. The main bottleneck appeared when interpreting the obtained rules since we deal with a massive number of rules where some of them are not contributing to the final understanding of new insights from data. We leave an open window for future research for this task, where we need to develop a new visualization method.

V. CONCLUSIONS

The goal of this paper was twofold. Firstly, we wanted to study how different ARM approaches to assist smart agriculture and if any specific approaches are based on NARM. Secondly, we wanted to set up a straightforward hardware environment for measuring soil parameters using cheap sensors utilized on Raspberry PI devices. We also wanted to develop a novel NARM method that can operate with timeseries data along with this hardware environment. A new algorithm called TS-NARM has been developed and evaluated on our dataset created on the data produced by soil and temperature sensors. Initial experiments suggest that there is a possibility of tailoring the basic NARM algorithm for dealing with time series data and widening the application to agriculture and other areas. Therefore, there is considerable room for improvement in the future also from the algorithmic perspective, e.g., how to mine rules where each day is an entity.

ACKNOWLEDGEMENT

This research has been partially supported by the project PID2020-115454GB-C21595 of the Spanish Ministry of Science and Innovation (MICINN).

REFERENCES

- Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases*, *VLDB*, volume 1215, pages 487–499. Citeseer, 1994.
- [2] Janez Brest, Sao Greiner, Borko Boskovic, Marjan Mernik, and Viljem Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE transactions on evolutionary computation*, 10(6):646–657, 2006.
- [3] Carlos Eduardo P Cerri, Gerd Sparovek, Martial Bernoux, Willian E Easterling, Jerry M Melillo, and Carlos Clemente Cerri. Tropical agriculture and global warming: impacts and mitigation options. *Scientia Agricola*, 64(1):83–99, 2007.
- [4] Bhagirath Singh Chauhan, Gulshan Mahajan, Ramanjit Kaur Randhawa, Harpreet Singh, Manjit S Kang, et al. Global warming and its possible impact on agriculture in india. *Advances in agronomy*, 123:65–121, 2014.
- [5] Sally Jo Cunningham and Geoffrey Holmes. Developing innovative applications in agriculture using data mining. In *The proceedings of the Southeast Asia regional computer confederation conference*, pages 25–29. Citeseer, 1999.
- [6] Javier Del Ser, Eneko Osaba, Daniel Molina, Xin-She Yang, Sancho Salcedo-Sanz, David Camacho, Swagatam Das, Ponnuthurai N Suganthan, Carlos A Coello Coello, and Francisco Herrera. Bio-inspired computation: Where we stand and what's next. *Swarm and Evolutionary Computation*, 48:220–250, 2019.
- [7] A. E. Eiben and James E. Smith. Introduction to Evolutionary Computing. Springer Publishing Company, Incorporated, 2nd edition, 2015.
- [8] AM Ezhilazhahi and PTV Bhuvaneswari. Iot enabled plant soil moisture monitoring using wireless sensor networks. In 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), pages 345–349. IEEE, 2017.
- [9] Cheng Fan, Meiling Chen, Xinghua Wang, Jiayuan Wang, and Bufu Huang. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers* in Energy Research, 9:652801, 2021.

- [10] Iztok Fister, Andres Iglesias, Akemi Galvez, Javier Del Ser, and Eneko Osaba. Differential evolution for association rule mining using categorical and numerical attributes. In *International conference on intelligent data engineering and automated learning*, pages 79–88. Springer, 2018.
- [11] John H Holland. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992.
- [12] Yaoguang Hu, Zhengjie Guo, Jingqian Wen, and Jialin Han. Research on knowledge mining for agricultural machinery maintenance based on association rules. In 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), pages 885–890. IEEE, 2015.
- [13] RC Jisha, G Vignesh, and D Deekshit. Iot based water level monitoring and implementation on both agriculture and domestic areas. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), volume 1, pages 1119–1123. IEEE, 2019.
- [14] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- [15] James Kennedy and Russell Eberhart. Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, volume 4, pages 1942–1948. IEEE, 1995.
- [16] Farah Khan and Divakar Singh. Association rule mining in the field of agriculture: a survey. *International Journal of Scientific and Research Publications*, 329, 2014.
- [17] Toshiharu Kojiri, Toshio Hamaguchi, and Mariko Ode. Assessment of global warming impacts on water resources and ecology of a river basin in japan. *Journal of Hydro-environment Research*, 1(3-4):164–175, 2008.
- [18] Nataša Kunstelj, D Ţnidarčič, and Branko Šter. Using association rules mining for sweetpotato (ipomoea batatas l.) in slovenia: a case study. *Int. J. Food, Agric. and Environ*, 11:253–258, 2013.
- [19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [20] Mark P McCarthy, Martin J Best, and Richard A Betts. Climate change in cities due to global warming and urban effects. *Geophysical research letters*, 37(9), 2010.
- [21] Amir Molajou, Parsa Pouladi, and Abbas Afshar. Incorporating social system into water-food-energy nexus. *Water Resources Management*, 35(13):4561–4580, 2021.
- [22] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2017.
- [23] Rainer Storn and Kenneth Price. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.
- [24] Amiya Kumar Tripathy, J Adinarayana, and D Sudharsan. Geospatial data mining for agriculture pest management-a framework. In 2009 17th International Conference on Geoinformatics, pages 1–6. IEEE, 2009.
- [25] N Vignesh and DC Vinutha. Association rule data mining in agriculture– a review. In *International Conference On Computational Vision and Bio Inspired Computing*, pages 233–239. Springer, 2019.
- [26] Peter M Vitousek. Beyond global warming: ecology and global change. Ecology, 75(7):1861–1876, 1994.
- [27] Grega Vrbančič, Lucija Brezočnik, Uroš Mlakar, Dušan Fister, and Iztok Fister. Niapy: Python microframework for building nature-inspired algorithms. *Journal of Open Source Software*, 3(23):613, 2018.
- [28] Xin-She Yang. Nature-inspired metaheuristic algorithms. Luniver press, 2010.