Applying Machine Learning to Extract New Knowledge in Precision Agriculture Applications

Savvas Dimitriadis[1], Christos Goumopoulos[*1,2]
[1]Hellenic Open University, 16, Sahtouri Str, Patras, Hellas
[2]Research Academic Computer Technology Institute,
N. Kazantzaki, 26500 Rio Patras, Hellas
goumop@cti.gr, sabbasdim@datascience.gr

Abstract— We are considering a facet of precision agriculture that concentrates on plant-driven crop management. By monitoring soil, crop and climate in a field and providing a decision support system that is able to learn, it is possible to deliver treatments, such as irrigation, fertilizer and pesticide application, for specific parts of a field in real time and proactively. In this context, we have applied machine learning techniques to automatically extract new knowledge in the form of generalized decision rules towards the best administration of natural resources like water. The machine learning application model suggested in this paper is based on an inductive and iterative process of discovering knowledge on the basis of which, patterns and associations having arisen initially are re-examined to expand the pre-existing knowledge. The result of this study was the creation of an effective set of decision rules used to predict the plants' state and the prevention of unpleasant impacts from the water stress in plants.

***Index Terms*— data mining, decision rules, machine learning, precision agriculture, machine learning process model**

1. INTRODUCTION

The urgent need to increase farming production, especially on an increasingly smaller piece of land, as well as the reduction of consuming resources such as water and fertilizers with respect to the environment, makes the use of new techniques and methods a first priority. Precision Agriculture (PA) is a suite of management strategies, technologies and practices that can solve the above problems. PA is an application of technologies and principles using information to manage spatial and temporal variability in order to increase the effectiveness of the resources and minimize environmental degradation. In other words, it is nothing but "doing the right thing, at the right time, in the right place, in the right way" [1], [2]. To make decisions for the achievement of the above goals, a basic approach is to monitor the plants' state and their environment throughout the year, and then analyse and interpret the data collected.

Thanks to developments in the field of wireless sensor networks as well as miniaturization of sensor systems, new trends have emerged in the area of PA. Wireless networks allow the deployment of sensing systems and actuation mechanisms at a much finer level of granularity, and in a more automated implementation than has been possible before. Sensors and actuators can be used to precisely control the concentration of fertilizer in soil based on information gathered from the soil itself, the ambient temperature, and other environmental factors. Incorporating feedback into the system through the use of sensors, actuators, and decision-making algorithms will allow a more fine-grained analysis that could adjust flow rate and duration in a way that is informed by local conditions.

Given a framework that gathers the necessary data, the decision making to be performed requires knowledge extraction from these data. Towards this direction, we have applied techniques known as knowledge discovery or data mining. Data mining is the process of discovering previously unknown and potentially useful information from data [3]. Machine learning is one of the most important and useful data mining tools that can discover unknown regularities and patterns from data sets.

The methodology we followed was based on applying machine learning techniques for inducing domain models of agriculture applications, which incorporates the process of discovering knowledge (data mining) with the close collaboration of the domain expert (agriculturist) and the machine learning expert.

Specifically, different machine learning strategies were studied and compared to each other regarding the analysis of agricultural data. Given the complexity of the parameters to be monitored and controlled in an agricultural environment, coupled with the possible imprecision of the information delivered an iterative and explorative evaluation process was applied to all available data sets in order to select the most useful features so as to improve the intelligibility and accuracy of the results. Then the best machine learning algorithms were chosen for the automatic extraction of knowledge from the data in the form of decision rules.

Finally, after taking into account the specificity of the application domain, all the derived rules were checked for their performance (based on the acceptable evaluation measures from the machine learning perspective), the scientific validity-reliability of the knowledge they express and the usage they have for the resolution of the problem examined.

The remainder of this paper is organised as follows. Related work is discussed in section 2. In section 3 we describe the phases of the machine learning process model applied for the

* Corresponding author

978-0-7695-3323-0/08 $25.00 © 2008 IEEE
DOI 10.1109/PCI.2008.30

application domain concerned. Based on this process model machine learning experiments on real agricultural data are presented in section 4. We conclude our paper with a discussion on the practical results of the machine learning application and the evaluation of the process model.

2. Related work

A proactive computing model by looping sensor data with actuators through a decision-making layer and the deployment of the system in a precision agriculture application was presented in [4]. In the work presented here, we actually follow and extend this model with a learning capability based on machine-learning algorithms which are used for inducing new rules by analysing logged datasets. The datasets used for the machine learning experiments in this work originate from the European IST FET Open project PLANTS [5]. The aim of this project was to optimise the efficiency and productivity of plant growth by using an array of sensors positioned around the crop which detects subtle plant/environmental signals and uses these as the basis for precision applications of water, pesticides or fertilisers.

Artificial intelligence and especially machine learning have contributed to the creation of control systems in agriculture. Neural networks have already been used since the 1990s for the creation of "smart" irrigation scheduling in the greenhouse environment [6]. In [7] the authors identified non-linear relationships between plant water status and the textural features of pictorial information of the plant canopy by using a layered neural network.

New Zealand is a characteristic example of a traditionally agricultural country which uses machine learning applications in farming activities. In [8] several projects in which machine learning has been used to assist data analysis are reported, such as a search for rules describing culling decisions in dairy herds, the isolation of factors governing apple bruising, the detection of cows "in heat" based on data collected during milking, and the analysis of a survey of microcomputer use in dairy farming.

Considering the sensitivity of plants on the changing climatic conditions, weather imponderables, pests etc., a system must be flexible and quick responding. In [9], this complexity/uncertainty is overcome by using fuzzy controllers for the sophisticated control of agricultural systems. The control input can be, for instance, environment (relative humidity) and the control output can be fruit response (water loss and skin color) during product storage.

Genetic algorithms were used for control optimization through simulation in a crop producing greenhouse. The objective was to maximize profit while minimizing the expenses of heating, CO_2 and electricity [10].

A key difference between the approaches considered above and our approach concerning the control and management of crop resources is that the former use optimization techniques that are based on measurements of the parameters only and do not aim to comprehend the way plants can operate. This fact as well as the inadequate understanding of key features of plant physiology, such as water stress, growth and photosynthesis, hampers their ability to build suitable mechanistic models for plant-based environmental control. Furthermore, many of the existing models have not been validated, or it is too difficult and costly to do so.

The solution to the above problems suggested in this study, uses machine learning and, as a result, complements the existing plant monitoring system [4]. The system is not only based on the plants responses, but mainly on data analysis, knowledge extraction and prediction of expected future conditions. The application of machine learning in the application domain leads to the simplification of the knowledge discovery process from real data, while at the same time it increases the reliability of the control system by minimizing its complexity and construction cost. The rules produced are directly connected with the knowledge and experience of the system as well as with the operations of the plants.

3. Learning Model

A. Learning Scheme Selection

A key goal of this work is to extract knowledge from available agricultural data and induce models that are in a easy manageable form by the decision making system and understandable to anyone involved in the crop production process.

After studying various learning techniques such as genetic algorithms and neural networks, we concluded that they follow a 'black box' approach during their operation and that they lack in modeling the knowledge they produce in a form comprehensible to humans. Genetic algorithms are a powerful technique used for optimization control systems, in which decisions for actions are based on the reaction to the current measurements from plants and their environment. Neural networks can learn by transforming their internal structure rather than by registering properly represented knowledge. Therefore, although the above learning techniques can have satisfactory performance and produce learning rules, they were avoided because they were considered unpractical and unsuitable for the problem discussed. Standard rule learning algorithms can produce models directly from data with classification or/and prediction accuracy analogous to, for instance, neural networks, but which are more comprehensible to humans/domain experts, than a neural model, thus our preference to the adoption of this technique.

Furthermore, despite the fact that the aim of the study is to solve a state diagnosis problem of the plants and these problems are satisfactorily solved from expert systems, the use of machine learning was considered necessary because the task of extracting rules is highly demanding and requires the expert to be fully aware of the problem. In some cases, the expert cannot supply information that can be incorporated into the knowledge basis of the expert system. On the contrary, the use of machine learning and especially of classification techniques initially requires the experts' knowledge (in the form of the class to which the data belongs) while later these techniques generate a model to explain the old and new examples. This process allows existing rule sets to be easily updated over time if the plants' state and their environment

change.

B. Learning Process Model

The data sets used consist of pre-classified examples. The classification was done by an expert of the application domain. The expert classified the examples through a set of domain rules. Our goal was to prescribe and verify the pre-existing knowledge as well as to incorporate new, since the domain model was not completed, by using machine learning algorithms.

For this reason and taking also into account the need to model knowledge in a comprehensible form, we have chosen classification algorithms, which produce a classifier as a set of rules or decision tree that can be then exploited to predict the classification of new data cases and can insert new rules to the domain model.

Figure 1 illustrates, as a data flow diagram, the machine learning process model, which is an expanded form of the process model for machine learning application in agriculture that was created by the Waikato Environment for Knowledge Analysis (WEKA) group in University of Waikato in New Zealand [11]. Our key contribution is to complement the original model by suggesting practical criteria for the extraction of the knowledge produced (see Table 4 for a summary of the practical guide for evaluation data sets and decision rules).

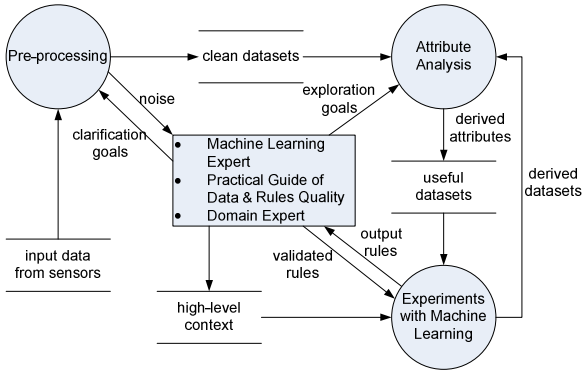

Fig. 1. Machine learning process model

The above process is iterative and requires the cooperation of the plant science domain expert and the data mining expert. In this study, the above cooperation was determined by the practical guide (see Table 4). The collaboration between the two experts is crucial to transform the raw data from the sensors into the final datasets to be used by the machine learning algorithms. In the pre-processing phase the context data may require tidying up, removing extraneous attributes, handling missing values, detecting erroneous values, etc. In the analysis phase, the domain expert will provide information about data semantics and legal relaxations or transformations that can be applied to the data, whereas the domain modeller will guide the process in order to improve the intelligibility and the precision of the results. Their interaction may reveal for example, that one or more attributes are irrelevant, or attributes may be manipulated mathematically to combine two or more columns into a single derived attribute. For the actual data mining process algorithms provided by the WEKA workbench are used (e.g., OneR, ZeroR, NaiveBayes, J48, DecisionStump, Nηge). ZeroR and NaiveBayes were used for generating a baseline performance metric that other learning schemes are compared with. Nηge is a nearest-neighbor method for generating rules [12].

Finally, in a post-processing phase the domain expert determines which part of the output rules is new useful knowledge to merit further exploration, and which part represent common knowledge.

All the derived rules were checked for their performance based on the acceptable evaluation measures from the machine learning perspective. In order to measure the error rate of the learning schemes we used the 10-fold cross-validation method. We also used as rates to compare the performance of algorithms the *overall success rate* (the number of correct classifications) and the *false positive rate (FP rate)*. The false positive (FP) rate is the proportion of examples which were classified as class X, but belong to a different class, among all examples which are not of class X.

4. Machine Learning Experiments

For the machine learning experiments datasets from the PLANTS project were used. The application data examined consider strawberry plants where the plant is controlling irrigation when the corresponding plant state is diagnosed. The prototype setup consists of an array of 96 plants placed in a glasshouse.

The plant and environmental parameters are: *ETR* (Electron Transport Rate by PAM meter[†]), *PAR* (light by PAR meter[‡]), *InfPAR* (inflection PAR – a derived attributed by combining ETR and PAR), *AmbC/PlantC* (Ambient/Plant leaf temperature by thermistors), *SM* (Soil Moisture by probe EC-10[§]), and the learning goals *Status* (HEALTHY/NOT HEALTHY), *HeatStress* (TRUE/FALSE), *DroughtStress* (TRUE/FALSE).

The *ETR* corresponds to the chlorophyll fluorescence parameter which forms the backbone of a feedback mechanism to determine the state of the photosynthetic rate of the plant and from this determine how productive the plant is under the current conditions. The *ETR* is calculated by combining the chlorophyll fluorescence and light *PAR* measurements. The *ETRxxx* correspond to the *ETR* value returned by the PAM meter for the saturated *PAR* value *xxx* (*xxx* corresponds to a sample light rate of 000, 050, 070, 105, 170, 270 and 425).

Status is the characterization of the plant general stress status. *DroughtStress* is the characterization of the plant water stress status in relation to the soil moisture levels. *HeatStress* can occur independently of water stress when the ambient environmental temperature gets very high and plant transpiration cannot maintain leaf cooling.

Data gathering as described above was performed for a

[†] Junior PAM, Gademann Instruments: http://www.gademann.com/

[‡] Skye SKP215 Quantum Sensor: http://www.alliance-technologies.net/meteo/PARTENAIRES/SKYE/skye.htm

[§] ECHO probe model EC-10: http://www.ech2o.com/specs.html.

period of several weeks corresponding to the early development stage of the crop and yielding a very large dataset. For the analysis purposes we have divided this dataset into two smaller. The first data set is called "ETR_Photosynthetic Activity" and includes 439 instances, 8 attributes and 1 class. The second set is called "eMultiPlant" and includes 1485 instances, 11 attributes and 3 classes. A segment of this dataset, used for running the machine learning algorithms with the WEKA workbench, is given in Table 1.

TABLE 1. A SEGMENT OF THE DATASET, USED FOR RUNNING MACHINE LEARNING ALGORITHMS WITH THE WEKA

ETR	Status	AmbC	Plant C	SM
387.1	OK	22.37	21.65	0.65
432.9	OK	22.37	21.65	0.64
412.8	OK	21.65	21.16	0.64
372.3	OK	21.65	20.92	0.64
463.8	OK	22.01	20.92	0.60
422.9	OK	22.01	20.67	0.60
315.8	OK	21.65	20.43	0.59
305.7	Not OK	21.28	20.43	0.59

The most relevant attributes were chosen in each data set and new important subsets arose. For example, in the "eMultiPlant" dataset we used the "GainRatioAttributeEval" attribute evaluator with the "Ranker" ranking method and 3 new important subsets arose, one for each classification goal (*Status, HeatStress and DroughtStress*). These data sets were evaluated by the domain expert as reliable who confirmed the correlation of the attributes with the prediction class.

Using WEKA we ran machine learning algorithms for all data sets and extracted a number of useful rules (Table 2). These rules take the form *IF antecedent condition(s) THEN consequent condition*, where antecedent condition is a conjunction of parameter-based tests and consequent condition is a category (e.g., a plant state such as healthy or not healthy). As we can see in Table 2, at least 2 rules were found to define each classifying variable of the plant. According to the domain expert, the rules that decide whether *DroughtStress* or *HeatStress* is true determine the request of irrigation.

TABLE 2. THE RULES SET DERIVED FOR RUNNING MACHINE LEARNING ALGORITHMS WITH THE WEKA

Rule id	Inferred Rule	Correctly Classified
1	**IF** *InfPAR* < 249 **THEN** *Status* is NOT HEALTHY	100%
2	**IF** *ETR425* <= 295 **THEN** *Status* is NOT HEALTHY	89%
3	**IF** *ETR270* <= 238 **THEN** *Status* is NOT HEALTHY	84%
4	**IF** *Δtemp* <= (-1) **THEN** *HeatStress* is TRUE	100%
5	**IF** (18.34 <= *AmbC* <= 28.5) **AND** (20.81 <= *PlantC* <=29.91) **AND** (3.39 <= *Δtemp* <= -1.3) **THEN** *HeatStress* is TRUE	100%
6	**IF** (*PlantC* > 24.42) and (*ETR50* <= 60) **THEN** *HeatStress* is TRUE	95%
7	**IF** *PlantC* > 27 **THEN** *HeatStress* is TRUE	96,5%
8	**IF** *SM* < 0.6 **THEN** *DroughtStress* is TRUE	98,2%
9	**IF** *InfPAR* < 243 **THEN** *DroughtStress* is TRUE	93,6%
10	**IF** *SM* <= 0.69 **THEN** *Status* is NOT HEALTHY	95%
11	**IF** *DroughtStress* == TRUE **THEN** *Status* is NOT HEALTHY	95%

The attribute *Δtemp* in rule 4 denotes temperature difference (*AmbC-PlantC*) and was derived in the Analysis phase.

5. DISCUSSION

The results of the methodology applied were satisfactory and showed the usefulness of machine learning to deal with real problems in the field of Precision Agriculture. The methodology was successful to extract new knowledge from the data and to induce it in an understandable and easily expandable form. This can positively contribute to the understanding and validation of the usefulness of control systems, which have predictable abilities, and can avert difficult conditions for the plants.

It is known that during the analysis stage, characteristics with lower informational value that do not play a significant role in the creation of the learning model, are usually removed from the dataset. Following another approach, we chose the best characteristics to improve the classification precision of the learning model without 'getting rid of' of the attributes initially not chosen as the best.

On the contrary, based on the suggested model and the practical guide (see Table 4) and by using the WEKA feature selection tools we analysed the relation between the attributes of lower informational value and the learning goals, and we finally created new data sets. The aforementioned methodology eliminated the danger of the appearance of highly complicated rules including many parameters often irrelevant to the classification class.

In addition, it offered a solution to a real problem appearing in crop control systems, i.e., the difficulty in gathering measurements of some main parameters due to possible damages caused to sensors. For instance, we see in Table 2 that the rules for the prediction of the *Status* can be based on 3 different parameters. Therefore, the inability to execute a rule due to the absence of its defining parameter can be easily and reliably dealt with the execution of a similar rule which uses another parameter.

Another interesting point is that the rules responded satisfactorily to the evaluated measures from the machine learning perspective. Table 3 shows the *overall success rate* and the *FP rate* for the classifiers of the *Status* learning goal. In the *FP Rate* column the values in parentheses refer to class NOT HEALTHY, whereas outside refer to class HEALTHY.

From Table 3 we can observe that the *FP rate* for class HEALTHY (i.e., the cases which incorrectly classified as HEALTHY while being NOT HEALTHY) takes the minimum value with the *DecisionStump* algorithm. On the other hand, the *OneR* algorithm gives the higher accuracy but with the higher *FP rate*. Considering it preferable to deal with a healthy state as being problematic than not to deal at all with an unhealthy plant state we chose the *DecisionStump* as better than another.

TABLE 3.OVERALL SUCCESS RATE AND THE FP RATE FOR STATUS

WEKA ML Classifier	**Correctly Classified**	**FP RATE**
NaïveBayes Simple	88.18 %	0.16 (0.11)
ZeroR	80.18 %	1 (0)
OneR	91.59 %	0.13 (0.07)
J48	89.41 %	0.09 (0.11)
DecisionStump	89.13 %	**0.05** (0.12)

The evaluation of the data quality and the validity of the produced knowledge were based on the practical criteria shown in Table 4. The necessity to use these control criteria arose from the research and the experiments performed for the sake of this study. These criteria can complement the existing machine learning application methodology on agricultural data, thus, increasing the reliability and effectiveness of crop control systems.

In conclusion, we should point out that through a number of experiments, new attributes were created, attribute subsets were chosen, thus creating new data sets, which ran in different learning models examining and evaluating different parameters each time. The attempt to evaluate the same states with different parameters was totally successful as far as this may be feasible. Therefore, the knowledge base of the system will be enriched with an expanded and precise set of rules. This fact further benefits the system, which can now support decision making even if a problem arises in the measurement gathering process of several basic parameters.

Although the inferred rules were reasoned by experts as rational to use, further work is needed towards the verification of the above estimation. This includes the execution of long term experimentations to verify repeatedly the assessments and to mask out possible errors or declinations. To gather and verify sensitivity to stress data, for example, it would be needed replicate populations of plants at the same growing stage but under the particular stress and one would also need to have complete control over all parameters (light, heat, humidity, fertilisation, irrigation, etc.); to gather the data there will be needed a large number of heterogeneous sensors and the experiment would have to be carried out repeatedly.

TABLE 4. PRACTICAL GUIDE FOR DATA SET AND DECISION RULES EVALUATION

1.	Evaluation of the usefulness of the plant and environmental parameters through application of machine learning techniques and the knowledge of a domain expert.
2.	Analysis of the states in which the plants can be found using all the available parameters.
3.	Classification of the majority of the attributes in new subsets according to their value and their connection with the learning goals.
4.	Evaluation of *Classification Precision* (percentage of successful rule covering in new examples).
5.	Cost evaluation of erroneous classifications.
6.	Evaluation of the *Knowledge Independence Rate* expressed by rules (low dependency on many attributes is desirable)
7.	Scientific evaluation of *Knowledge Validity* from an application domain expert.
8.	Evaluation of the *Transferability* and *Management* of the rules.

REFERENCES

[1] Srinivasan, A., Handbook of Precision Agriculture Principles and Applications, Haworth Press, 2006.

[2] Karydas, C., and Silleos, N., Precision Agriculture: Current state and Prospects in Greece. In: Proceedings of the 2nd Special Conference of Applications of Informatics to the Agricultural Sector, Hellenic Operational Research Association, Chania, Greece, October 2000.

[3] Piatesky-Shapiro, G., and Frawley, W., (Eds.), Knowledge Discovery in Databases, AAAI Press, 1991.

[4] Goumopoulos, C., Kameas, A., and Oflynn B., Proactive Agriculture: An Integrated Framework for Developing Distributed Hybrid Systems, in Proc. of the 4th International Conference on Ubiquitous Intelligence and Computing (UIC-07), Jadwiga Indulska et al. (Eds.), Springer-Verlag, pp 214-224, 2007.

[5] Cassells, A., Goumopoulos, C., Morrissey, A., and Tooke F., New crop of technology reveals plant health, ICT Results, [on-line] http://cordis.europa.eu/ictresults/index.cfm?section=news&Tpl=article&BrowsingType=Features&ID=81342, April 2006.

[6] Ehret, D.L., A. Lau, S. Bittman, W. Lin and T. Shelford, Automated monitoring of greenhouse crops. Agronomie 21: 403-414, 2001.

[7] Murase H., Honami N., Nishiura Y., A neural network estimation technique for plant water status using the texture features of pictorial data of plant canopy, Acta Hort. 399:255–262, 1995.

[8] Garner, S.R., Holmes G., McQueen R.J. and Witten I.H., Machine learning from agricultural databases: practice and experience, New Zealand J Computing 6 (1a), pp. 69-73, 1995.

[9] Morimoto T., Hashimoto Y., and Hoshi T., An intelligent control technique based on fuzzy controls, neural networks and genetic algorithms for greenhouse automation, in: Kozai T., Murase H. (Eds.), Proc. 3rd IFAC-CIGR Workshop, Artificial intelligence in agriculture, Chiba, Japan, April 1998, pp. 61-66, 1998.

[10] Krink, T., Ursem, R., K., and Filipic, B., Evolutionary Algorithms in Control Optimization: The Greenhouse Problem, In: Proc. of the 3rd Genetic and Evolutionary Computation Conference, p. 440-447, 2001.

[11] Garner, S.R., et al., Applying a machine learning workbench: experience with agricultural databases, In: Proc. Machine Learning in Practice Workshop, Machine Learning Conference, pp. 14-21, 1995.

[12] Witten, I.H., and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2005.