Application of Data Mining in automatic description of yield behavior in agricultural areas

M.G. Canteri[1], *B.C. Ávila*[2], *E.L. dos Santos*[2], *M.K. Sanches*[1], *D. Kovalechyn*[1], *J.P.Molin*[3], *L.M. Gimenez*[3]

Comentário:

Abstract

Data mining is the non-trivial extraction of implicit knowledge in databases which aims to retrieve useful and new information in a high level of abstraction. The advent of Precision Farming generates databases which, because of their size and complexity, are not efficiently analyzed by traditional methods. The present work aims to test if Data Mining routines are capable to determine the behavior of the yield of a crop as a function of physical-chemical soil properties, in order to allow correction of low yield. Databases were used as object of work, where yield is the meta-attribute and the physical-chemical soil properties are the predictive attributes, obtained through data acquisition on field, in areas of 400 m2. The meta-attribute was obtained by way of a yield map generated by precision agriculture equipment. The data set, with 2388 records were mined using the Decision Tree technique and all values were discerned into two levels. As result, rules were generated – finite sets of pairs attribute-value – describing models relating yield and physical-chemical soil properties. The confidence of rules was evaluated automatically, making possible the selection of the most qualified ones. By the analysis of the rules by human experts, it was determined that is possible to use the models to determine the behavior of the yield of a crop as a function of physical-chemical soil properties. The developed tool is still without processing acceleration techniques and without techniques of refinement of quality of discovery knowledge, what recalls expectations that the results can be quite improved.

Keywords: Decision Support Systems, Knowledge-based Systems, Precision Farming

Introduction

Data mining works with the discovery of hidden knowledge, unexpected patterns and new rules from large databases (Adriaans & Zantinge, 1996). Data mining is the non-trivial extraction of implicit knowledge in databases which aims to retrieve useful and new information in a high level of abstraction. The goal of a data mining process is to discovery data in order to recover information not easily recovered from database by other methods, like SQL language. That information could be used in a decision process (Ávila, 1998).

There are a great variety of data mining methods and the decision tree is one of the most used. The decision tree classify examples of a finite number of classes

[1]Lab. InfoAgro, Dep. de Informática, Univ. Estadual de Ponta Grossa, CEP 84030-000, Ponta Grossa, PR, Brazil, mgcanter@uepg.br. [2] Department of Informatics, Pontificia Universidade Católica do Paraná, CEP 80215-901, Curitiba, PR, Brazil. [3] Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, CEP 13418-900, Piracicaba, SP, Brazil.

(Janikow, 1998). The nodes of tree are represented by attributes names, the connection are represented by possible value to the attribute and the leaf with the different classes. An object is classified following the path from root tree to the leaf, in agreement with the satisfied connections (attribute value).

The advent of Precision Farming generates databases that because of their size and complexity are not efficiently analyzed by traditional methods. Precision farming or precision agriculture has potential for maintaining or improving crop yields while reducing required chemical inputs. Its able to achieve its goals by considering a field as a group of sub-areas than as one homogenous area. Using a Global Positioning System (GPS) with a Geographical Information System (GIS), field data can accurately be collected. Although high technology equipment is available to apply chemicals (fertilizer and defensives), management decision-making remains a significant problem its on-farm use (Hayes & Privette, 1998).

The purpose of this study was to test if Data Mining routines are capable to determine the behavior of the yield of a crop as a function of physical-chemical soil properties, in order to allow correction of low yield.

Material and Methods

An database was used as object of work and the data mining process was used as a descriptive tool, in order to describe the meta-attribute behavior. The database had information about soybean yield (meta-attribute) and physical-chemical soil properties (predictive attributes), collected in the region of Campos Novos, SP, Brazil. The data were obtained in the following way:

- The area was divided into 2388 rectangular cells of same dimension;
- Every cells had information about yield and physical-chemical soil properties;
- The information were grouped and collected using GPS devices;

The meta-attribute was obtained by way of a yield map generated by precision agriculture equipment. The predictive attributes were achieved directly from cells or obtained by interpolate calculation from neighborhood areas. There were 17 predictive attributes: *pH* (related to soil acidity), *ctc* (related to soil ion exchange capacity), *v* (saturation), *h_al* (hydrogen plus aluminum), *Ca* (calcium), *Mg* (magnesium), *Mn* (manganese), *P* (phosphorus), *K* (potassium), *Bo* (boron), *Zn* (zinc), *Cu* (copper), *Fe* (iron), *m_o_*(organic matter), *sand*, *silt* and *clay*.

The present data mining study used the Decision Tree technique, based in the C45 algorithm (Quinlan, 1993), using a Prolog routine. The knowledge discovery techniques require discrete data, but the database had continuous attributes. So it was used a process to transform discrete to continuous data. The transformation was oriented to meta-attribute yield and was one of the process most time costly, in spite of use only two classes. It means that a attribute had a value above or below of a calculated threshold. The meta-attribute threshold was fixed at 2000 kg/ha and the threshold to the predictive attributes were calculate by the software routine.

Also, there was a pre process and a after process, relative to application of C45 algorithm. The pre-process stages transform the database in a adequate

representation to the mining process, and the after process stages were applied to refine the knowledge in order to present only the valuable information.

As result, rules were generated – finite sets of pairs attribute-value – describing models relating yield and physical-chemical soil properties. The confidence of rules was evaluated automatically, making possible the selection of the most qualified ones.

Results and Discussion

The discovered knowledge was represented by *if-then* rules. The group of conditions denoted after *if* condition, for a cell, represented the yield level attained after *then* sentence with the marked confidence. If the inverse order was followed we can found the soil profiles that determine a certain yield level, what makes possible a human analysis to find the more influential substances, positive or negatively.

For instance, the results generate by prolog routines after processing are a series of rules like:
IF *clay < 19.815* ***THEN*** *yield >= 2000* ***CONFIDENCE*** *= 1.0*
It means that for the tested cells, when clay level was lower than 19.815 the yield was equal or higher than 2000 kg/ha. This example have only one condition but the results presented rules with up to 5 conditions influencing the yield. It is desirable to have a rule with lower number of conditions and a high confidence value.

Rules for yield higher or equal than 2000 kg/ha

The Table 1 presents rules generated by the tested routine for yield higher or equal to 2000 kg/ha. It was selected only rules with confidence equal to 1.0 (100 %). Analyzing the results we found coherent and fascinating rules and others that need to be tested under field condition.

A coherent rule is represented by rule number 6 (Table 1) where a yield upper to 2000 kg/ha its represented by a high level of *P*, high level of *K*, low level of *Zn* and high level of *Bo*. But, the rule number 7 indicate that in a sand soil (*sand >=80.725*) the levels of *Zn* and *Bo* are inverse.

It was detected that the *P* levels always was higher than threshold in the rules for *yield >= 2000*, what is coherent with real world. But, analyzing the Table 2 (*yield < 2000*) we also found a larger number of rules with *P >= 38.265*. It indicates that there are some relationship among tested elements and its dangerous to look just for an isolate attribute inside the rules.

Rules for yield lower than 2000 kg/ha

The Table 2 presents rules for *yield < 2000*. The rule number 1 presents a typical coherent rule, there are a lower yield if *pH* was low and organic matter (*m_o_*) was low.

Analyzing the number of each element appeared inside the rules and if it was upper or lower than threshold we found results presented at Table 3. Inside the 20 rules, 45% of then (9 rules) had the *pH* element, and 7 presented conditions with *pH* lower than threshold. When we use the elements with 30% or more of occurrence and analyze if they are upper or lower than threshold we can to set up a new rule. The rule formed by this way for *yield < 2000* was:
pH < 5.97 and v < 70.86 and Zn < 1.221 and Bo > 0.1596 and silt < 4.354

Time of Processing

The time of processing was evaluated with 2 computers. A Pentium 100 MHz take approximately 180 hours to generate the rules and a Pentium III 800 Mhz reduced the time of processing to approximately 11 hours. That time do not include the pre processing and the after processing routines.

Forward

This first work was oriented to simplicity and velocity instead of accuracy. It was a first step work, seeking for forward routines and improvements. It was used only two levels of yield, the use of a higher number of yield levels require a watchful study to avoid small number of conditions to higher and upper levels.

The developed routine deserve tuning and new evaluation about the used techniques. For instance, could be tested more levels of the elements (not only two) and others techniques instead of Decision Tree (Avila, 1998) in order to improve the quality of generated information and to reduce the time of processing. Its also necessary to appeal to a human expert to define the threshold levels.

Others improvements are required to pre-processing and after-processing steps, in order to improve the quality of data to be mined and to refine the discovered knowledge.

After definition of appropriated technique and improvements application will be design a automatic tool (software) to analyze the databases improving the automation. The software should have a simple interface and a good presentation of discovered knowledge in order to facilitate the interpretation of rules.

The analysis of the rules by human experts indicated that is possible to use the models to determine the behavior of the yield of a crop as a function of physical-chemical soil properties. The developed tool is still without processing acceleration techniques and without techniques of refinement of quality of discovery knowledge, what recalls expectations that the results can be quite improved.

We can conclude that data mining is a promising tool to analyze data from precision farming. Using the appropriate technique we could transform a regular production field in a experimental field to find relationships among elements and yield.

References

Adriaans, P. and D. Zantinge 1996. Data Mining. Addison-Wesley, England, 158 pp.

Ávila, B.C. 1998. Data Mining, Escola de Informática da SBC - Regional Sul, Blumenau. pp.87-106.

Hayes, J. C. and C. V. Privette. 1998. Potential benefits of precision farming. In: Zazueta, F.S. and J. Xin. Proceedings of the 7th International Conference, Orlando, Florida, pp. 70-78.

Janikow, C.Z. 1998. Fuzzy Decision Trees: Issues and Methods, IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, n.1, pp.1-14.

Quinlan, J.R. 1993. C45: programs for machine learning, Morgan Kaufman Publishers, San Mateo, California.

Tables

Rule	*if*	*and*	*and*	*and*	*and*
1	*clay < 19.815*				
2	*pH >= 5.97*	*P >= 38.265*	*Cu >= 0.7476*	*Bo < 0.1596*	
3	*pH >= 5.97*	*P >= 38.265*	*m_o_ < 19.705*	*Zn < 1.221*	
4	*pH < 5.97*	*v >= 70.86*	*K < 1.542*	*Zn >= 1.221*	*silt >= 4.354*
5	*P >= 38.265*	*ctc < 62.94*	*Zn < 1.221*	*Bo >= 0.1596*	
6	*P >= 38.265*	*K >= 1.542*	*Zn < 1.221*	*Bo >= 0.1596*	
7	*P >= 38.265*	*K >= 1.542*	*Zn >= 1.221*	*Bo < 0.1596*	*sand >= 80.725*
8	*ctc >= 62.94*	*Mn >= 9.5045*			

Table 1: Rules for yield higher or equal than 2000 kg/ha with confidence equal to 1,0

Rule	*if*	*and*	*and*	*and*	*and*
1	*pH < 5.97*	*m_o_ < 19.705*			
2	*pH < 5.97*	*ctc < 62.94*	*Ca >= 28.865*		
3	*pH < 5.97*	*v >= 70.86*	*Zn < 1.221*	*silt < 4.354*	
4	*pH < 5.97*	*Bo >= 0.1596*	*silt < 4.354*		
5	*pH < 5.97*	*Bo >= 0.1596*	*Ca < 28.865*		
6	*pH < 5.97*	*Bo >= 0.1596*	*Cu < 0.7476*	*sand < 80.725*	
7	*pH < 5.97*	*Bo >= 0.1596*	*Zn >= 1.221*	*Cu < 0.7476*	
8	*pH >= 5.97*	*Fe >= 18.15*	*silt >= 4.354*		
9	*pH >= 5.97*	*K < 1.542*	*Ca >= 28.865*	*silt < 4.354*	
10	*ctc < 62.94*	*m_o_ >= 19.705*	*silte < 4.354*		
11	*ctc < 62.94*	*P >= 38.265*	*K < 1.542*	*Fe >= 18.15*	
12	*ctc < 62.94*	*P >= 38.265*	*v < 70.86*	*Zn < 1.221*	
13	*ctc >= 62.94*	*Mn >= 9.5045*			
14	*v < 70.86*	*m_o_ < 19.705*	*Zn < 1.221*		
15	*v < 70.86*	*Bo >= 0.1596*	*silt < 4.354*		
16	*v < 70.86*	*Bo >= 0.1596*	*P < 38.265*	*Zn >= 1.221*	*Sand >= 80.725*
17	*v < 70.86*	*Bo < 0.1596*	*Fe < 18.15*		
18	*v < 70.86*	*P >= 38.265*	*Ca < 28.865*	*Zn < 1.221*	
19	*v >= 70.86*	*m_o_ >= 19.705*	*K < 1.542*	*Zn < 1.221*	*silt < 4.354*
20	*K < 1.542*	*Cu < 0.7476*	*Zn < 1.221*	*sand < 80.725*	

Table 2: Rules for yield lower than 2000 kg/ha with confidence equal to 1,0

predictive attribute	threshold	Number of rules		Occurrence (%)
		> threshold	< threshold	
pH	5.97	2	7	45
ctc	62.94	1	4	25
v	70.86	2	6	40
m.o.	19.75	2	1	15
P	38.26	3	1	20
K	1.54	0	3	15
Ca	28.86	2	2	20
Bo	0.159	6	1	35
Zn	1.22	2	6	40
Cu	0.747	0	3	15
Fe	18.15	2	1	15
sand	80.72	0	2	10
silt	4.354	1	5	30
clay	19.81	0	0	0

Table 3: Number of times that rule appear in conditions higher or lower than threshold for yield lower than 2000 kg/ha.