Automatic Classification Technique in an Environmental Unit Recognition: A Case Study

L.H.A. Rodrigues[a] and A.C.B. Munari[b]

[a] *Faculdade de Engenharia Agricola – FEAGRI, Universidade Estadual de Campinas – UNICAMP, Caixa Postal 6011, 13083-970 – Campinas, SP, Brazil (lique@agr.unicamp.br)*

[b] *Faculdade Prudente de Moraes, R. Prof. José Benedicto Gonçalves, 309, 13306-230 – Itu, SP, Brazil (acmunari@splicenet.com.br)*

Abstract: Environmental unit recognition is a multidisciplinary task that gathers elements of several knowledge fields, such as geology, biology, geography and ecology. This paper aims to present the benefits of using a tool that automatically generates classification rules in a case study of environmental unit recognition. Data from a previous academic research, which led to an M.Sc. Thesis, was used in this research. Results obtained by human experts analysis and by using automatic classification techniques were similar in many aspects. Comparison of these two approaches showed that this tool provides considerable advantage over human expert analysis. It greatly simplifies the development of deterministic aspects of the research, rendering it automated. This allows human experts to spend most of their time in analysis of aspects where human intervention is absolutely necessary, such as tasks involving creativity and judgements.

Keywords: Induction Rules; Automatic Classification; Environmental Analysis; Data mining

1. INTRODUCTION

Many environmental research projects are currently developed all over the world with different costs and infrastructure requirements, according to their objectives and region where these projects are applied. Methodologies used in these projects usually involve phases such as (i) mapping and photo interpretation of pictures of the region, (ii) gathering material and field observations, (iii) laboratory analysis and (iv) interpretation of results. The first three tasks are generally systematic and objective, with well-defined formal methods, while the latter is usually a time-consuming investigation effort, strongly dependent on expert sensitivity and experience. As a consequence, experts require a considerable amount of time to elaborate conclusions and validate their hypothesis. This makes it very difficult to analyze massive amounts of data leading to a better comprehension of the relationship among several components of the system. Consequently, contributions to facilitate data analysis, with the objective of acquisition of evidences from available data are welcome, since they are instrumental in accelerating the final analysis process. They can also improve the precision and quality of conclusions in some circumstances. Studies in Statistics and Computer Science have been performed to adequately deal with these kind of problems [Johnson, 1998; Han, 1995; Michalski et al., 1998], particularly to those of knowledge management resulting in efficient techniques to solve them systematically.

This paper aims to analyze the necessary possibilities and conditions concerning the use of a tool that induces automatic classification rules in an environmental unit recognition research. The adoption of such a tool allows the automation of a very important part of the analysis phase referred to as the exploratory process. It allows the expert to concentrate with more emphasis on aspects naturally more dependent on human intervention, such as those related to creativity or judgement. Classification rules are suitable for this category of problems for several reasons: (i) classification is often present in any research; (ii) automatic classification rules make the comprehension easy and does not require human intervention at the beginning of the process; (iii) nowadays, there are efficient and easy-to-use tools to accomplish the classification task. Therefore, in the present research, we verify the viability of using data classification tools based on decision trees in order to accelerate the analysis process, resulting in interpretation and validation of the model in an environmental study. A more detailed description of the procedures and results can be found in Munari [2001], with an extensive reference list of papers.

2. MATERIAL AND METHODOLOGY

2. 1 Introduction

We evaluated the rule induction technique using data from an environmental research previously performed without this technique. The M.Sc. Dissertation titled "Geomorphologic Analysis and Vegetation Spatial Distribution at the Picinguaba Coast (Ubatuba, São Paulo, Brazil) [Garcia, 1995] was chosen since it provided relevant data as well as it was a good example of human expertise requirement in several knowledge fields.

In our research we selected the software See5. This software has been developed based on the classic algorithms for rule induction ID3 and C4.5 by Ross Quinlan [Quinlan, 1993].

2.2 Brief Description of the Environmental Research

The main objective of the research was to verify how suitable it is to use detailed geomorphologic mapping in environmental unit recognition and identify relationships between environment and vegetation. Typically, this is a multidisciplinary task, present in many recent environmental studies, which integrates several knowledge fields such as geology, biology, geography and ecology. The research also made it viable to gather knowledge from a complex environment, organize it and make it available for future proposals to the handling of environmental management and preservation problems.

The research was executed in a wild coast of São Paulo State, called "Picinguaba Coast". A detailed map of the region was generated and used to identify and find main common characteristics of several natural regions, called "Genetically Homogeneous Surfaces" (GHS).

GHSs were initially determined based on analysis of aerial photos of the region followed by field observations. Based on these group definitions, 22 different points were selected, where soil samples were collected. Physical composition analysis and photo interpretation techniques were combined to check the GHS characteristics.

Soil samples were collected from pre-determined layers: every 10 cm from 10 cm down to 50 cm; every 20 cm from 50 cm down to the groundwater level. There were a total of 160 soil samples. Laboratory analysis were considered under three perspectives:

- Macroscopic: made with dry material; evaluation of color, texture and mineralogy
- Granulation: determination of proportions of Total Sand (TS), Silt (S) and Clay (C); later TS was decomposed in 5 granulation classes, according to Wentworth Scale.
- Chemical: determination of proportions of Carbon, Organic Matter (OM), pH, Nitrogen, Phosphorus, Sodium, Aluminum, Calcium, etc.

Analysis results were compared with GHSs classification in order to find any relationship between physical soil composition and GHSs that could characterize them.

Ten GHSs were defined for the region (GHS I to GHS X).

According to the author, comparisons between the classification defined by geomorphologic mapping and physical composition analysis of soil was shown to have a high relation for each GHS, specially when macroscopic and granulation characteristics were considered. There were only a few samples with chemical analysis since evidences were considered strong enough to eliminate their use, particularly their high cost.

2.3 Brief Description of the See5

See5 is a computer program that reads database files (called training set and test set) and generates decision trees by induction, according to previously determined categories. A decision tree represents a classification rule in a hierarchical way, as shown in Figure 1.

See5 Release 1.12 for Windows was used to induce rules in this paper.

Once data is formatted there are two main steps:

- Classification generator: when rules characteristics are adjusted by several parameters that affect accuracy and representation rules and data interpretation.

- Interactive use with new data: when non-classified data are used as input (one by one) and the predicted class is given as the program output together with a coefficient that indicates the certainty of that prediction.

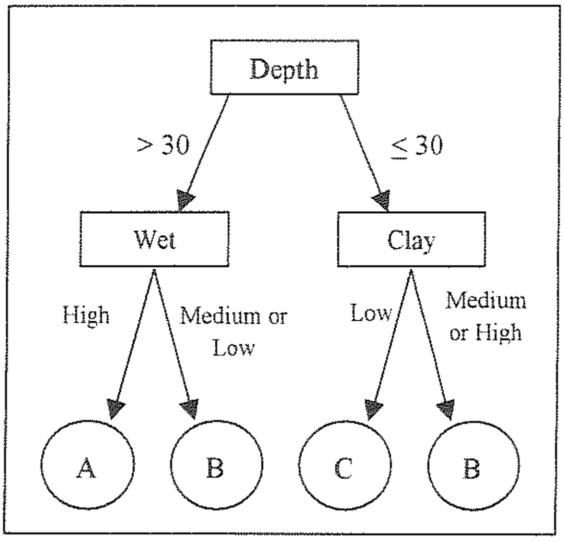

Figure 1. An example of a decision tree.

See5 generates decision trees from Information Theory concepts. Initially, the program minimizes the system entropy via splitting the training set into the most homogeneous subsets [Quinlan, 1993].

See5 provides many features that have been used in this research such as:

- Post pruning : sub-trees are discarded when error rate is predicted to be high;
- Clustering : it allows automatic grouping of discrete attributes that generalize concepts; this tends to generate compact decision trees;
- Rules: IF-THEN rule sets can be obtained from a training set; these are frequently more legible and compact than branches of the corresponding trees; the results are similar to the example below:

```
Rule 3:  (3/1, lift 1.2)
         Wet = Medium
         Depth > 30
         →  class B  [ 0.600 ]
```

Rule number is automatically generated; there is an information about the number of instances of the training set used to generate the rule and the relationship between this number and the confidence index; a set of attributes (rule conditions); a classification presented next to the symbol "→" that corresponds to the Right Hand Side of the rule and a value in the interval [0, 1] that indicates the rule confidence. The interpretation of this rule is:

```
Rule name: Rule 3
Number of instances:  3
Instances incorrectly classified:  1
Rule:     IF Wet = "Medium"  AND  Deep > 30
          THEN class = "B"
Confidence Index:  60 %
No. of instances  X  Confidence Index:  1.2
```

Since the classifier is generated See5 reports a summary of the parameters that were effectively used, a description of the decision tree and the rule sets generated and a brief compilation of the errors obtained with the training set (and the test set if applicable).

2.4 Strategy

During first steps data are collected and adequately organized to be processed by data mining tools. Typically this task corresponds to data management (to integrate data, to eliminate redundancy, etc.). Then, data is converted into a specific format required by the tool. See5, for example, uses text files (ASCII): a file with general data (metadata) has extension ".names" and a file with values to be computed has extension ".data". Result analysis is typically a task for human experts that can use graphs and other visual resources.

Data from the original environmental research consisted of approximately 160 samples, each one with 30 attributes. They describe the 27 different places of data collection (macroscopic, granulation and chemical aspects) indicating respective GHSs. With these data sets we could identify classification rules that allow the association of each sample with its GHS, a typical data mining task. Data used in this research have a relatively high dimensionality (30 attributes) and sample number that is not proportional to its dimensionality (160 samples). Additional difficulty occurred since soil samples were collected from different depths. This brings vertical relationships in each collection point that makes data modeling even more difficult. Another important aspect to be considered was the different number of collection points of soil samples for each GHS (in some cases with only one or two); for operational and financial reasons, it was impossible to obtain new data samples from these GHSs. Although such constraints may affect conclusions about relevant relationships obtained in this research, they can still be useful to indicate that this procedure can be successfully applied to similar research projects.

2.5 Brief Description of the Experimental Work

We compared results from automatic classification tools in environmental problems against the ones obtained from a human expert. This helped us evaluate the suitability of those tools for this kind of problem. We selected two experiments and applied the technique with data from the experimental research. In Experiment 1 we used data collected from layer depths between 10 cm and 20 cm. In Experiment 2, we used data from layers between 10 cm and 40 cm, where effects of vertical relationships are reduced. Our experiments have been labeled (1a, 1b, 2a, 2b, etc.), each letter corresponding to the attribute sets taken into account for the specific classification; this allows us to understand the effect of each attribute set on the final result. Experiment 1a, for example, is the first version of the Experiment 1 and takes into account only granulation to determine GHSs. Experiment 1b takes into account macroscopic data.

Each experiment is described in details and results obtained with each one is subsequently presented and compared to the original environmental research [Garcia, 1995], in order to verify the relationship between the two classifications.

2.5.1. Experiment 1 - Classification by granulation and macroscopic analysis

This procedure aimed to gather elements to verify the relationship between the observed characters in granulation and macroscopic aspects of each soil sample and each respective GHS. This is similar to what has been accomplished in the reference research. This verification was performed by using an induction classifier and a training set where relative proportions of sand, clay and silt are specified, as well as GHS code corresponding to each sample. Results are then shown as decision trees and rule sets (antecedent and subsequent). This allows an expert to understand criteria used to build them. A more complete version of this experiment also takes into account other important attributes of macroscopic analysis such as Selection Level, Texture and Mineral Composition of soil samples.

In an ideal situation, each classifier should then be applied to a new sample data (called test set) with the same structure as the training set, in order to estimate errors in the classification. However, in our case, the amount of data was not enough to apply this procedure. Hence, in our case, error rate is the one obtained with the training set despite its low reliability, since it tends to be lower than a real error rate. Automatic classification results are then compared to the one obtained from the reference work, which allows us to verify the suitability and quality of criteria used for the induction.

2.5.2. Experiment 2 - Classification by granulation and chemical and macroscopic analysis

This procedure aimed to verify whether there is any chemical composition influence on GHS characteristics. In this case, the strategy was to generate a training set similar to the Experiment 1, but including only chemical composition data for each sample soil with its corresponding GHS. A new classifier was obtained and evaluated similarly to the first one.

Another classifier was obtained by using a training set where granulation, macroscopic and chemical aspects were considered simultaneously. This allows to compare this classification (theoretically more complete than Experiment 1) with those obtained by considering only one or two attributes.

3. RESULTS AND DISCUSSION

Classifiers were generated for all experiments, each one induced by its respective training set. We present below a brief discussion of results obtained from Experiment 2b, with soil samples collected between 10 and 40 cm. A complete discussion of this and the other experiments can be found in Munari [2001].

Decision trees generated by See5 are graphically presented in order to make the classifier structure more comprehensible.

In Experiment 2b, 36 soil samples with 21 attributes each were taken into account (only 20 attributes were actually considered to induce rules). Data included granulation and chemical analysis obtained between 10 and 40 cm.

Each GHS is represented in the tree by a leaf, according to Figure 2. Results showed that attributes related to granulation are much more important than those related to chemical analysis (only Mg and Na actually affected part of the tree). Hence, Experiment 2b results were very similar to those obtained when only granulation was considered at the same depth (Experiment 2a).

Two subtrees were generated. The first one, where Coarse Sand (CS) is significantly present (more than 60%), corresponds to soil with a high diversity and with a predominant presence of sandy materials. GHSs VI, VII, VIII and X are present in this group. They are relatively old and this is characterized by the presence of grains with size between 0.5 and 1.0 mm. In addition, in GHS VI there is a significant presence of Very Fine Sand (VFS) that makes it quite different from GHSs VII, VIII and X, where a low presence of VFS is observed. Similarly, we observed a low presence of Total Sand (TS) in GHS X, what makes it particularly different from the others in this group. Finally, a relatively high proportion of Fine Sand (FS) was observed in GHS VIII. In the second subtree we observed that low proportions (but significant) of silt characterized GHSs III and IV, which are distinguished by the level of Mg (much higher in GHS III). Soil samples from GHS I presented low level of silt (less than 1.4 %, when the average was 11.0 %) and high level of Na (over 1.34 ppm, when the average was 0.95 ppm). Regions with low levels of Na belong to GHSs II and V, which are distinguished by the TS level (higher in GHS II).

When results obtained with the automatic classifier were compared with classification by the human expert in the original research, an important aspect to be observed is that granulation is the predominant attribute for the classification.

It is also observed that GHSs VI, VII, VIII and X are characterized by high diversity of sand, typically coarse. GHSs I, II, III, IV and V have a higher uniformity in granulation (fine and very fine sand soils with low or no clay or silt). This makes the classification based only on granulation to be more difficult. In these cases the importance of chemical composition increases since it facilitates the finding of evidences that characterize each GHS. Results have shown that GHSs I and II have a very high TS and VFS levels that do not allow to differentiate soil samples from each one based only on this attribute. However, Sodium (Na) level in GHS I is significantly higher than in GHS II. GHS III is also difficult to be characterized based only on granulation, but when Magnesium (Mg) level is considered it helps to differentiate GHSs III and IV, the first one with significantly higher Mg level.

Results have shown that there are two more general groups of GHSs in all experiments: the first one (GHSs I, II, III, IV and V), with some cases difficult to classify based only on granulation; in the other group, with more diversified lands (GHSs VI, VII, VIII, IX and X), coarse sand soils are more predominant. These two large groups are present in almost all

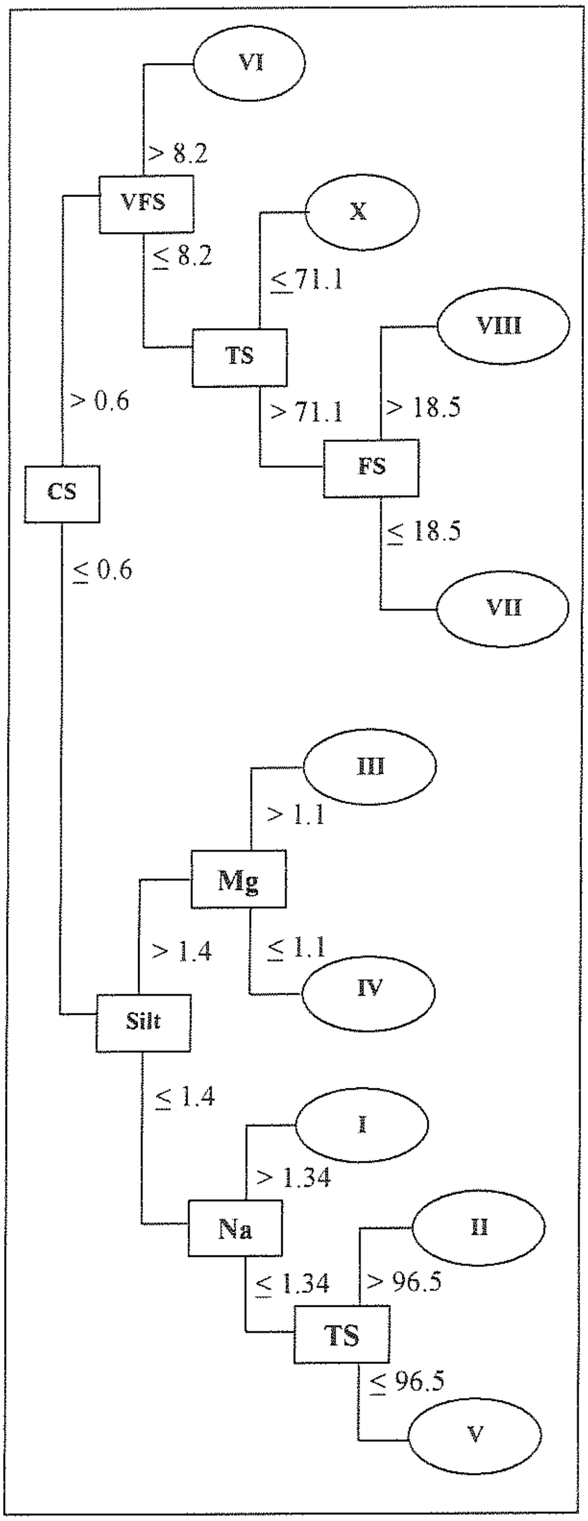

Figure 2: Decision tree induced for Experiment 2b.

experiments. Results on Experiment 2a were not as accurate as the others due to the low quantity of soil samples.

Some important changes could be observed in the training set when soil samples from different depths were considered altogether. It may be only the consequence of low quantity of soil samples in each experiment, but it can also be an indication of structural changes in the composition of lands due to the increasing depth.

Chemical data have not significantly affected the classification probably due to the reduced amount of samples. Induced classifiers based only on those attributes were not good and the combination of them with granulation data led to decision trees where only Mg was important. It seems to be worth verifying this chemical element more carefully since it affected classification of GHS III in all experiments. This region was always more difficult to be classified based only on the other attributes and presence of Mg have shown to be a specific characteristic of it.

4. CONCLUSION

The first general conclusion is that material collection for this sort of experiment must take into account its utilization for data mining. This precaution is necessary in order to allow a better use of any data mining tool, reducing time spent with pre-processing of data.

The second general conclusion is that rule induction classifiers allow accelerating the data analysis phase, since it helps to quickly extract potentially interesting logical structures from considerable amount of data, where human intervention is very difficult or unfeasible.

Furthermore, classifier representations with decision trees and production rules seem to be intuitive to be understood and actually used even by researchers with no knowledge of machine-learning techniques. However, such tools should be seen as tools to give support to the decision-maker and not to make the decision. Participation of human experts is essential to interpret results and recommend some eventual fine adjustments, according to the objectives of the research. Data mining tools only present clues raised from available data. The human expert must interpret them, evaluate them and decide whether they are relevant or not. This judgement skill is also necessary to evaluate how much database limitations impact on conclusions.

Concerning the present paper, some specific relevant conclusions were found. First of all, results should be viewed with reserve since the rules were based on a low number of soil samples. In addition, some factors such as soil samples from different depth levels and absence of data from some soil samples (particularly concerning chemical analysis), typical of these sort of researches due to high costs to collect, enhanced these difficulties. Despite these aspects, results have shown that this approach was efficient. They were very similar to those the human expert found with the same low quantity of data. Results obtained with the automatic classification by induction also have shown that granulation and macroscopic data seem to be enough to classify GHSs.

Finally, some other complex analysis can be included in the study. A practical example is to include data of variety of orchids in each GHS that can characterize biodiversity level. This can be determined by the number of species of orchids found in each region, which is highly related to the biodiversity of a region.

5. REFERENCES

Garcia, J.P.M., Geomorphologic analysis and vegetation spatial distribution at the Picinguaba Coast (Ubatuba, São Paulo, Brazil), M.Sc. Dissertation. State University of São Paulo, Brazil, 176 p., 1995 *(in Portuguese).*

Han, J., From database systems to knowledge-based systems: an evolutionary approach. *In:* XI International Conference on Data Engineering, Taipei, Taiwan, Conference Tutorial. 1995.

Johnson, R. A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 4th ed, Upper Saddle River (NJ), Prentice-Hall, 816 pp., 1998.

Michalski, R.S., I. Bratko and M. Kubat (ed.), *Machine Learning and Data Mining: Methods and Application*, Baffins Lane (UK), John Wiley & Sons, 456 pp., 1998.

Munari, A.C.B., Use of automatic classification techniques in environmental analysis: a case study, M.Sc. Dissertation, State University of Campinas (Unicamp), Brazil, 140 p., 2001 *(in Portuguese).*

Quinlan, J.R., C4.5: *Programs for Machine Learning*, San Mateo (CA), Morgan Kaufmann, 302 pp., 1993.