Induction of Landtype Classification Rules from GIS Data

L. Karl Branting[1], William A. Reiners[2], and Hongyan Wang[3]

Abstract

The feasibility of inducing classification rules for Landtype Associations (LTAs) from instances of human-expert classifications was tested by evaluating the accuracy of 3 rule-induction algorithms on data drawn from a GIS coverage of Southeast Wyoming. In 10-fold cross-validation tests, the accuracy of rules using precipitation, vegetation, geology, elevation, slope, and aspect as features achieved over 87% accuracy. Adding position as a feature increased accuracy to over 95%. Pruning rule sets to increase comprehensibility caused only a slight decrease in accuracy, particularly for the most accurate induction algorithm, RIPPER. The evaluation indicates that human expert LTA classification rules can be effectively induced from examples and applied to large GIS coverages.

1 Introduction

Development and application of landscape taxonomies is an important task in landscape ecology. Landscape taxonomies are expressed in Landtype Associations (LTAs), groupings of landtypes based on similarities in geomorphic process, geological rock types, soil complexes, stream types, lakes, wetlands, and vegetation communities (ECOMAP 1993).

The growing availability of GIS coverages means that LTA classifications are no longer restricted to regions that can be surveyed on the ground. Instead, LTA classifications can be extended to broad GIS coverages. However, analysis of large quantities of GIS data is a difficult and error-prone process. Development of computer tools to help automate the induction and application of LTA rules would therefore be an important step in the application of LTAs to large-scale GIS coverages.

This paper describes an experimental evaluation of the feasibility of using machine-learning algorithms to induce LTA rules from examples of human expert LTAs. The evaluation was intended to address the following issues:

1. How can the LTA task be formulated as a machine-learning problem?
2. What is the relative accuracy of various rule-induction algorithms when applied to this problem?
3. Can the comprehensibility of large induced rule sets be improved by shrinking their size without compromising accuracy?

The rule-induction algorithms were trained with examples drawn from a survey of Wyoming classified by Reiners and Thurston (1997). The ability of the induced rules to predict the classification of other points in the survey area was then tested. As set forth below, the results of the experimental evaluation provide preliminary confirmation that LTA rules having acceptable levels of accuracy and comprehensibility can be induced from examples and then applied to large-scale GIS data.

2. Landtype Associations in Southeast Wyoming

ECOMAP (1993) is a system of landtype mapping adopted by several land-managing agencies in the U.S. including the USDA Forest Service, USDOI Bureau of Land Management and National Park Service. These agencies are the principle managers of Federal land

[1] Department of Computer Science, University of Wyoming , Laramie, WY 82071, USA, email: karl@uwyo.edu.
[2] Department of Botany, University of Wyoming, Laramie, WY 82071, USA, email: reiners@uwyo.edu.
[3] IBM Global Services, Boulder, CO, USA, email: davewang@us.ibm.com.

in Wyoming. ECOMAP is a derivative of Bailey's ecoregional approach (Bailey 1995, 1996) which focuses on terrestrial aspects of land environments including climate, prevailing vegetation and land form. It stands in contrast with the Omernick system (Omernick 1987) which is based on fluvial drainage systems.

As defined in ECOMAP, a landscape is:

> a heterogeneous land area composed of a cluster of interacting ecosystems that is repeated in a similar form throughout; and can be viewed at one time from one place.

Landtype Associations are large-scale landscape regions reflecting:

> similarities in geomorphic process, geological rock types, soil complexes, stream types, lakes, wetlands, and series, subseries, or plant association vegetation communities.

Partitioning landscapes into LTAs is important for ecologists and natural-resources managers because LTAs are the fundamental conceptual units for ecological modeling at the landscape scale.

The delineation of LTAs in southeast Wyoming by Reiners and Thurston (1997) was performed manually. The principal objective of the delineation was to produce a digital map of landscape ecosystems for management purposes, such as decision support, regional planning, and cumulative impact assessment.

Reiners and Thurston's LTA delineation was based on six $60m^2$ ARC/INFO coverages of southeast Wyoming. These coverages represent the following features:

1. Precipitation
2. Vegetation
3. Geology
4. Elevation
5. Slope
6. Aspect

Precipitation is the mean annual precipitation in millimeters. Precipitation values were generated by a model described in Daly, Neilson, and Phillips (1994). *Vegetation* is the land cover type occupying the largest area within each $60m^2$ cell. There are 41 vegetation types. The *geology* coverage identifies which of 333 geological features characterizes each cell.

Elevation is the height in meters, derived from digital elevation model (DEM) data from a U.S. Geological Survey (1987). *Slope*, in degrees, is derived from the elevation data. Finally, *aspect* is the illumination of each cell given the topology of the landscape and a fixed illumination source.

Using these ARC/INFO coverages, Reiners and Thurston created the following set of LTA classes based on geomorphic properties:

1. Mountains
2. Isolated mountain
3. Granite hills
4. Channeled hills
5. Dissected plateau
6. Footslope
7. Breaks
8. Mesa(s)
9. Scoria hill
10. Recessional escarpment
11. Multiple cuesta and valley complex
12. Single cuesta
13. Glacial till and outwash hills
14. Rolling plains
15. Rolling plains and alluvial valleys
16. Lakes/reservoir

Reiners and Thurston then used these LTA classes to delineate Southeastern Wyoming based on the six GIS coverages described above. In forming the delineation, they were guided by the criterion that each LTA should be recognizable by land users and managers in the field and should be sufficiently different from surrounding terrain that they would consider its management differently. The resulting LTA delineation, which is represented in an additional GIS coverage, is shown in Figure 1.

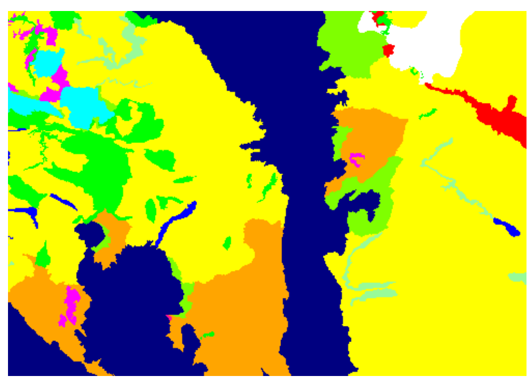

Figure 1. Landtype Associations for Southeast Wyoming derived manually by Reiners and Thurston (1997).

3. Inducing LTA Rules
3.1 Rule-Induction Algorithms

The machine-learning and data-mining communities have developed a large number of algorithms for *supervised concept learning*, the task of inducing a classifier from a set of classified instances (Langley 1996). In general, supervised concept learning algorithms perform hill-climbing search through a hypothesis space of possible classifiers, guided by a heuristic evaluation function that combines the classifier's accuracy on a training set with some measure of parsimony, e.g., fewer assumptions or smaller hypothesis size. Supervised concept-learning algorithms differ in (1) the nature of the hypothesis space, e.g., decision trees or rules, (2) the strategy used to search the hypothesis space, e.g., bi-directional search, divide-and-conquer, and (2) the heuristic function that guides search.

In our initial experiment, we used three induction algorithms: C4.5 (Quinlan 1993); CN2 (Clark and Niblett 1989); and Ripper (Cohen 1995). These algorithms were selected because each embodies a different inductive bias, and because all three can produce rules that can be understood and evaluated by human experts.

Three experiments were performed to test the feasibility of inducing LTA rules for GIS data from classified instances.

3.2 Experiment 1: Can LTA Rules Be Learned?

Reiners and Thurston were guided by their intuitions as ecologists in making LTA delineations. The purpose of our first experiment was to test our hypothesis that the judgments of ecologists, typified by the delineation judgments of Reiners and Thurston, are governed by rules that can be induced from examples. Inducing such rules would be useful because they could permit the process of LTA delineation to be partially or entirely automated.

In the first experiment, the predictive accuracy of the 3 induction algorithms was compared in 10-fold cross-validation on 20,000 samples uniformly drawn from the 10,799,169 cells in the LTA coverage derived by Reiners and Thurston, including only those portions of Southeast Wyoming for which values for all six coverages were available. (Although the learning methods can handle missing values, we found accuracy to be severely compromised). The error rates are summarized in Table 1.

Table 1. Mean error rates in 10-fold cross-validation tests.

Algorithms	Ripper	CN2	C4.5
Error Rate	13.75	13.96	14.61

These results indicate that a moderately accurate model of human LTA criteria can be induced from examples. Ripper was slightly more accurate than the alternative learning algorithms on this data set.

Figure 2 shows the LTA's predicted by the rule set induced by Ripper. The dark rectangular areas are missing one or more feature values.

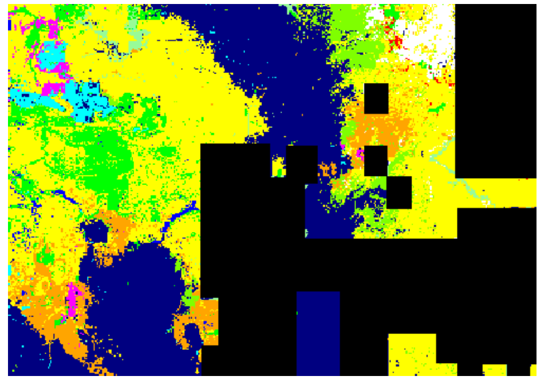

Figure 2. LTAs predicted by the rule set induced by Ripper.

The cells for which the induced LTA rules conflict with the LTA assignments by Thurston and Reiners are shown in gray in Figure 3.

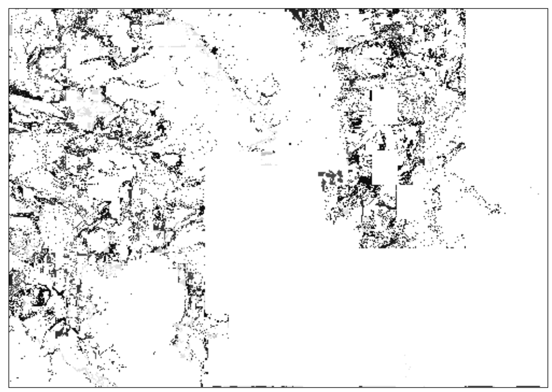

Figure 3. Cells on which the rule set induced by Ripper conflicts with Reiners andThurston's LTAs.

3.3 Experiment 2: Do Human Experts Use Spatial, as Well as Featural, Information?

Several alternative hypotheses could explain why the accuracy of the induced rule set was substantially less than 100%. One possibility is that the human experts are inconsistent in their judgments and therefore that no consistent model is possible. A second hypothesis is that the inductive biases of the learning algorithms are ill-suited for the human experts' LTA categories. A third hypothesis is that the set of features used to define the instances is incomplete.

Reiners and Thurston based their LTA delineation on the same six GIS coverages used to generate the data sets used by the learning algorithms. However, by viewing the coverages as a whole, Reiners and Thurston had access to an additional form of information: the large-scale boundaries between regions with similar feature values. This information about the boundaries between regions is lost when training instance are sampled randomly without regard to the spatial relationships among the samples.

To test whether the loss of spatial information was responsible for the errors in the induced rule sets, we repeated the 10-fold cross validation experiment with Ripper using the x,y coordinates of each cell as additional features. The results of the experiment, set forth in Table 2, indicate that adding spatial data substantially increases the predictive accuracy of the induced rule set.

Data Set	Original data set	x,y coordinates added
Error Rate	13.75	4.43

Table 2. Error rates for rule sets induced by Ripper with and without x,y coordinates.

The results of experiment 2, which showed that adding x,y coordinate can lead to over 96% accuracy, suggest that models of human expert LTA delineation must include not just local features but also the aggregation of these features into large-scale units.

3.4 Experiment 3: Pruning Rule Sets

The rule sets induced by Ripper and the other rule induction algorithms tested were often quite large, typically several hundred rules. Such large rule sets are generally quite difficulty to understand, even by domain experts. Comprehensibility is an important attribute of induced rule sets because more comprehensible rules are easier to validity and more likely to be accepted by experts.

All other things being equal, rule-set comprehensibility is an inverse function of rule-set size. Thus, the smaller of two equally accurate rule sets should be preferred. In practice, however, the usefulness of a rule set may be a function both of its accuracy and of its comprehensibility. Thus, some trade-off between rule-set size and accuracy may be acceptable, i.e., a large gain in comprehensibility may compensate for a small loss of accuracy.

Experiment 3 investigated the trade-off between accuracy and comprehensibility. Unfortunately, there is no general theory of rule comprehensibility. We therefore used rule-set size as an approximate measure of comprehensibility. For each of the three induction algorithms—Ripper, CN2, and C4.5—we used the standard rule-pruning mechanism for that learning method, systematically varying the parameter controlling the extent of pruning (code weight, exception weight, and internal threshold, respectively). For each rule set size, the accuracy of each rule set was tested in 10-fold cross validation on the same dataset as was used in Experiment 2 (6 features plus x,y coordinates).

The results, which are set forth in Figure 4, indicate little loss of accuracy in any of the algorithms from shrinking the rule set from its original size of almost 400 by a factor of 2. When the induced rule set falls below 150, accuracy begins to fall off more rapidly. However, for highly pruned rule sets, Ripper appears to perform somewhat better than CN2 or C4.5.

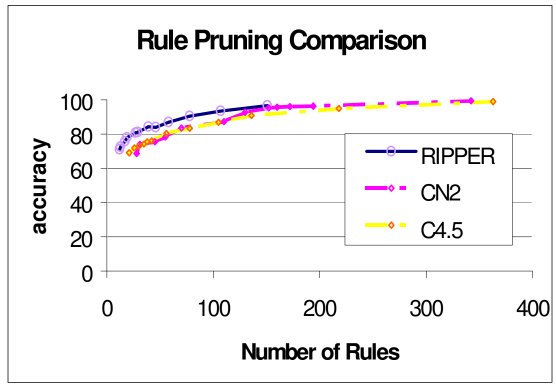

Figure 4. Accuracy of three learning algorithms as a function of rule-set size.

3.5 Example Rule Set

An example of a highly pruned rule set generated by Ripper is shown in Figure 5. Two typical rules in this set are the following:

```
Lake/Reservoir:-
vegetation  =  open water (141/29).

Granite Hills:-
precipitation< = 273, geology = wg
(46/16).
```

This first rule means that if the vegetation type is "open water" then the landtype is `Lake/Reservoir`. The "`(141/29)`" following the rule means that the rule correctly classified 141 instances, but classified 29 other instances incorrectly. Similarly, the second rule means that if the annual precipitation value is less than 273 millimeter and geology type is wg, then the landtype is `Granite Hills`. The second rule correctly classified 46 instances, but incorrectly classified 16 others.

The rules are order-invariant, that is, they can be applied in any order, unlike a decision list. However, there is a default rule: if no other rule applies, then the default category, `rolling plains`, is applied.

```
Final hypothesis is:

Lake/Reservoir:-
vegetation = open water (141/29).

Granite Hills:-
precipitation< = 273, geology = wg
(46/16).

Major River Valley:-
vegetation = open water, geology = Qa,
precipitation> = 326, precipitation< =
362, shaded_relief> = 999 (53/17).

Major River Valley:-
elevation < = 1327, precipitation< = 347
(32/7).

Isolated Mountain:-
geology = wg, precipitation< = 415,
elevation > = 2070, elevation < = 2499
(83/4).

Isolated Mountain:-
geology = wg,   vegetation = v42009,
precipitation> = 559 (28/13).

Isolated Mountain:-
geology = PM, elevation > = 2339 (26/1).

Dissected Plateau:-
elevation < = 1650, geology = p&h
(137/0).

Dissected Plateau:-
elevation < = 1697,   vegetation =
Ponderosa pine, elevation < = 1564,
precipitation> = 340 (40/8).

Footslopes:-
elevation < = 1942, geology = Tmu,
precipitation< = 363, precipitation> =
344 (175/39).

Multiple Cuesta and Valley Complex:-
precipitation< = 317, geology = Tha,
elevation < = 2200 (210/101).

Rolling Plains and Alluvial Valleys:-
precipitation< = 339, elevation > = 2206,
precipitation> = 303 (709/229).

Rolling Plains and Alluvial Valleys:-
geology = Qa, elevation > = 2174,
precipitation< = 346 (227/93).

Rolling Plains and Alluvial Valleys:-
vegetation = irrigated crops,
shaded_relief< = 72, elevation > = 2071
(175/67).

Rolling Plains and Alluvial Valleys:-
precipitation< = 359, geology = Qt,
precipitation> = 313 (124/64).

Rolling Plains and Alluvial Valleys:-
precipitation< = 359,   vegetation =
irrigated crops, geology = Qs,
shaded_relief< = 34 (67/0).

Mountain:-
precipitation> = 443, elevation > = 2396,
precipitation< = 456 (2120/89).

Mountain:-
precipitation> = 430, geology = wg
(491/8).

Mountain:-
geology = Wgn (642/39).

Mountain:-
precipitation> = 440, geology = Ys
(192/0).

Mountain:-
precipitation> = 406, geology = Yla
(162/0).

default
Rolling Plains(9505/3791).

Train error rate: 23.08% +/- 0.30%
(20000 datapoints)
Hypothesis size: 21 rules, 79 conditions
Learning time: 115.84 sec
```

Figure 5. A highly pruned rule set generated by RIPPER from 20,000 training instances.

4. Summary

The three experiments illustrated how current rule-induction algorithms can be successfully used to help automate the formalization of LTA classification rules from examples of human-expert LTAs. Experiment 2 indicates that landscape ecologists may be guided by spatial information, such as the shape and continuity of boundaries separating regions with distinct geo-morphology. Addition of x,y coordinates, while demonstrating the importance of spatial information, is not a feasible approach for inducing rule sets for regions distinct from the area from which the training instances have been sampled. Instead, some approach to integrating clustering with classification is needed for exploiting spatial information of this type.

Experiment 3 showed that rule-set accuracy falls off fairly slowly with rule-set size. This suggests that when small rule sets are important for comprehensi-bility and validation, rule sets can be extensively pruned without compromising acceptable accuracy.

Acknowledgments

This research was supported in part by a grant from the Andrew W. Mellon Foundation.

References

Bailey, R. G., 1995. *Description of the Ecoregions of the United States.* (2nd ed., 1st ed. 1980). Misc. Publ.

No. 1391, Washington, D.C. U.S. Forest Service. 108 pp with separate map at 1:7,500,000.

Bailey, R. G., 1996. *Ecosystem Geography*. Springer-Verlag. 216 pp.

Clark P. and Niblett T., 1989. The CN2 Induction Algorithm, *Machine Learning* 3 (4), 261-283.

Cohen, W., 1995. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning,* Lake Tahoe, California.

Daly, C., Nielson, R., and Phillips, D., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Met.* 33:140-158.

ECOMAP, 1993. *National hierarchical framework of ecological units*. Unpublished administrative paper. Washington, DC. U.S. Department of Agriculture, Forest Service. 20p.

Langley P., 1996. *Elements of Machine Learning*, Morgan Kaugmann, San Francisco, CA.

Omernik, J. M., 1987. *Ecoregions of the Conterminous United States.* Map (scale 1:7,500,000). Annals of the Association of American Geographers.

Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning,* Morgan Kaufmann.

Reiners W. and Thurston R., 1997. *Delineations of Landtype Associations For Southeast Wyoming,* Final Report.