Discovering Communicable Scientific Knowledge from Spatio-Temporal Data

Mark Schwabacher   MARK.SCHWABACHER@ARC.NASA.GOV

NASA Ames Research Center, Computational Sciences Division, MS 269-1, Moffett Field, CA 94035 USA

Pat Langley   LANGLEY@ISLE.ORG

Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306 USA

Abstract

This paper describes how we used regression rules to improve upon a result previously published in the Earth science literature. In such a scientific application of machine learning, it is crucially important for the learned models to be *understandable* and *communicable*. We recount how we selected a learning algorithm to maximize communicability, and then describe two visualization techniques that we developed to aid in understanding the model by exploiting the spatial nature of the data. We also report how evaluating the learned models across time let us discover an error in the data.

1. Introduction and Motivation

Many recent applications of machine learning have focused on commercial data, often driven by corporate desires to better predict consumer behavior. Yet scientific applications of machine learning remain equally important, and they can provide technological challenges not present in commercial domains. In particular, scientists must be able to *communicate* their results to others in the same field, which leads them to agree on some common formalism for representing knowledge in that field. This need places constraints on the representations and learning algorithms that we can utilize in aiding scientists' understanding of data.

Moreover, some scientific domains have characteristics that introduce both challenges and opportunities for researchers in machine learning. For example, data from the Earth sciences typically involve variation over both space and time, in addition to more standard predictive variables. The spatial character of these data suggests the use of visualization in both understanding the discovered knowledge and identifying where it falls short. The observations' temporal nature holds opportunities for detecting developmental trends, but it also raises the specter of calibration errors, which can occur gradually or when new instruments are introduced.

In this paper, we explore these general issues by presenting the lessons we learned while applying machine learning to a specific Earth science problem: the prediction of Normalized Difference Vegetation Index (NDVI) from predictive variables like precipitation and temperature. We begin by reviewing the scientific problem, including the variables and data, and proposing regression learning as a natural formulation. After this, we discuss our selection of regression rules to represent learned knowledge as consistent with existing NDVI models, along with our selection of Quinlan's Cubist (Rulequest, 2001) to generate them. Next we compare the results we obtained in this manner with models from the Earth science literature, showing that Cubist produces significantly more accurate models with little increase in complexity.

Although this improved predictive accuracy is good news from an Earth science perspective, it comes as little surprise to those with a background in machine learning. However, in our efforts to communicate the discovered knowledge to our Earth science collaborators, we have also developed two novel approaches to visualizing this knowledge spatially, which we report in some detail. Moreover, evaluation across different years has revealed an error in the data, which we have since corrected. We discuss some broader issues that these experiences raise and propose some general approaches for dealing with them in other spatial and temporal domains. In closing, we also review related work on scientific data analysis in this setting and propose directions for future research.

2. Monitoring and Analysis of Earth Ecosystem Data

The latest generation of Earth-observing satellites is producing unprecedented amounts and types of data about the Earth's biosphere. Combined with readings from ground sources, these data hold promise for testing existing scientific models of the Earth's biosphere and for improving them. Such enhanced models would let us make more accurate predictions about the effect of human activities on our planet's surface and atmosphere.

One such satellite is the NOAA (National Oceanic and Atmospheric Administration) Advanced Very High Resolution Radiometer (AVHRR). This satellite has two channels which measure different parts of the electromagnetic spectrum. The first channel is in a part of the spectrum where chlorophyll absorbs most of the incoming radiation. The second channel is in a part of the spectrum where spongy mesophyll leaf structure reflects most of the light. The difference between the two channels is used to form the Normalized Difference Vegetation Index (NDVI), which is correlated with various global vegetation parameters. Earth scientists have found that NDVI is useful for various kinds of modeling, including estimating net ecosystem carbon flux. A limitation of using NDVI in such models is that they can only be used for the limited set of years during which NDVI values are available from the AVHRR satellite. Climate-based prediction of NDVI is therefore important for studies of past and future biosphere states.

Potter and Brooks (1998) used multiple linear regression analysis to model maximum annual NDVI[1] as a function of four climate variables and their logarithms:

- Annual Moisture Index (AMI)

- Chilling Degree Days (CDD)

- Growing Degree Days (GDD)

- Total Annual Precipitation (PPTTOT)

These climate indexes were calculated from various ground-based sources, including the World Surface Station Climatology at the National Center for Atmospheric Research. Potter and Brooks interpolated

[1]They obtained similar results when modeling minimum annual NDVI. We chose to use maximum annual NDVI as a starting point for our research, and all of the results in this paper refer to this variable.

the data, as necessary, to put all of the NDVI and climate data into one degree grids. That is, they formed a 360×180 grid for each variable, where each grid cell represents one degree of latitude and one degree of longitude, so that each grid covers the entire Earth. They used data from 1984 to calibrate their model. Potter and Brooks decided, based on their knowledge of Earth science, to fit NDVI to these climate variables by using a piecewise linear model with two pieces. They split the data into two sets of points: the warmer locations (those with $GDD \geq 3000$), and the cooler locations (those with $GDD < 3000$). They then used multiple linear regression to fit a different linear model to each set. They obtained correlation coefficients (r values) of 0.87 on the first set and 0.85 on the second set, which formed the basis of a publication in the Earth science literature (Potter & Brooks, 1998).

3. Problem Formulation and Learning Algorithm Selection

When we began our collaboration with Potter and his team, we decided that one of the first things we would do would be to try to use machine learning to improve upon their NDVI results. The research team had already formulated this problem as a regression task, and in order to preserve communicability, we chose to keep this formulation, rather than discretizing the data so that we could use a more conventional machine learning algorithm. We therefore needed to select a *regression learning* algorithm — that is, one in which the outputs are continuous values, rather than discrete classes.

In selecting a learning algorithm, we were interested not only in improving the correlation coefficient, but also in ensuring that the learned models would be both understandable by the scientists and communicable to other scientists in the field. Since Potter and Brooks' previously published results involved a piecewise linear model that used an inequality constraint on a variable to separate the pieces, we felt it would be beneficial to select a learning algorithm that produces models of the same form. Fortunately, Potter and Brooks' model falls within the class of models known as *regression rules* in the machine learning community (Weiss & Indurkhya, 1993). A regression rule model consists of a set of linear models and a set of inequality "cuts" on the variables to select among the individual linear models, yielding a piecewise linear model. To induce such rules, we selected Cubist, a commercial product from Rulequest Research (2001), which has evolved out of earlier work with C4.5 (Quinlan, 1993) and M5 (Quinlan, 1992).

MIN. RULE COVER	No. RULES	r
1%	41	0.91
5%	12	0.90
10%	7	0.89
15%	4	0.88
20%	3	0.86
25%	2	0.85
100%	1	0.84

4. First Results

We ran Cubist using the same data sets that Potter and Brooks had used to build their model, but instead of making the cuts in the piecewise linear model based on knowledge of Earth science, we let Cubist decide where to make the cuts based on the data. The results exceeded our expectations. Cubist produced a correlation coefficient of 0.91 (using ten-fold cross-validation), which was a substantial improvement over the 0.86 correlation coefficient obtained in Potter and Brooks' earlier work. Potter and his team were pleased with the 0.91 correlation coefficient, but when we showed them the 41 rules produced by Cubist, they had difficulty interpreting them. Some of the rules clearly did not make sense, and were probably a result of Cubist overfitting the data. More importantly, the large number of rules — some 41 as compared with two in the earlier work — was simply overwhelming.
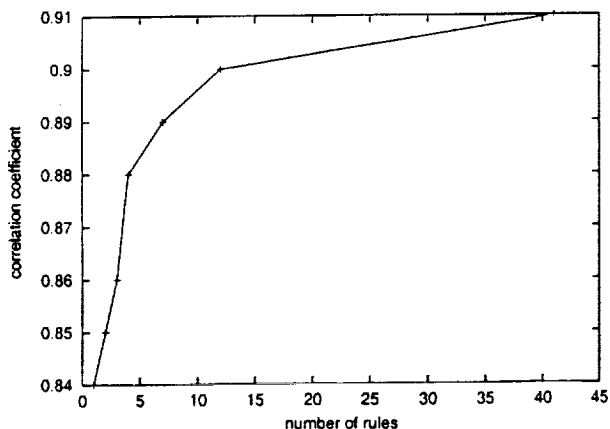

Figure 1. The number of rules in the Cubist model and the correlation coefficient for several different values of the minimum rule cover parameter.

The first step we took in response to this understandability problem was to change the parameters to Cubist so that it would produce fewer rules. One of these

Table 2. The two rules produced by Cubist when the minimum rule cover parameter is set to 25%.

```
Rule 1:

if
   ppttot <= 25.457
then
   fasmax = -3.22465 + 7.07 ppttot + 0.0521 cdd
            - 84 ami + 0.4 ln(ppttot) + 0.0001 gdd

Rule 2:

if
   ppttot > 25.457
then
   fasmax = 386.327 + 316 ami + 0.0294 gdd
            - 0.99 ppttot + 0.2 ln(ppttot)
```

parameters specifies the minimum percentage of the training data that must be covered by each rule. The default value of 1% produced 41 rules. We experimented with different values of this parameter between 1% and 100%; the results appear in Table 1 and Figure 1. Using a model with only one rule — that is, using conventional multiple linear regression analysis — results in a correlation coefficient of 0.84, whereas adding rules gradually improves accuracy. Interestingly, when using two rules, Cubist split the data on a different variable than the one the Earth scientists selected. Potter and Brooks split the data on GDD (essentially temperature), while Cubist instead chose precipitation, which produced a very similar correlation coefficient (0.85 versus 0.86). The two-rule model produced by Cubist is shown in Table 2.

In machine learning there is frequently a tradeoff between accuracy and understandability. In this case, we are able to move along the tradeoff curve by adjusting Cubists' minimum rule cover parameter. Figure 1 illustrates this tradeoff by plotting the number of rules and the correlation coefficient produced by Cubist for each value of the minimum rule cover parameter in Table 1. We believe that generally a model with fewer rules is easier to understand, so the figure essentially plots accuracy against understandability. A useful feature for future machine learning algorithms would be the ability to directly specify the maximum number of rules in the model as a parameter to the learning algorithm.[2] We used trial and error to select values for the minimum rule cover parameter that produced the number of rules we wanted for understandability reasons.

[2]After reviewing a draft of this paper, Ross Quinlan decided to implement this feature in a future version of Cubist.
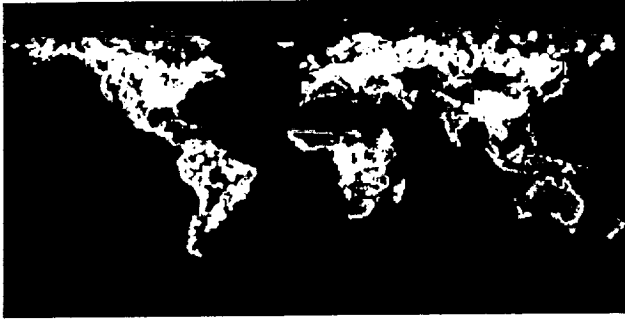
Figure 2. Map showing which Cubist rules are active across the globe.

Figure 3. Map showing the errors of the Cubist prediction of NDVI across the globe.

5. Visualization of Spatial Models

Reducing the number of rules in the model by modifying Cubists' parameters made the model more understandable, but to further understand the rules, we and the Earth scientists decided to plot which rules were active where. In Figure 2, the black areas represent portions of the globe that were excluded from the model because they are covered with water or ice, or because there was insufficient ground-based data available.[3] The white areas are regions in which more than one rule in the model applied. (In these cases, Cubist uses the average of all applicable rules.) The gray areas represent regions in which only one rule applies; the six shades of gray correspond to the six rules. (We normally use different colors for the different rules, but resorted to different shades of gray for these proceedings.)

Potter and his team found this map very interesting, because one can see many of the Earth's major topographical and climatic features. The map provides valuable clues as to the scientific significance of each rule. This type of visualization could be used whenever the learning task involves spatial data and the learned model is easily broken up into discrete pieces that are applicable in different places, such as rules in Cubist or leaves in a decision tree.

A second visualization tool that we developed shows the error of the Cubist predictions across the globe. In Figure 3, black represents either zero error or insufficient data, white represents the largest error, and shades of gray represent intermediate error levels. From this map, it is possible to see that the Cubist model has large errors in Alaska and Siberia, which is consistent with our collaborators' belief that the quality of the data in the polar regions is poor. Such a map

---

[3]After excluding these areas, we were left with 13,498 points that were covered by the model.

can be used to better understand the types of places in which the model works well and those in which it works poorly. This understanding in turn may suggest ways to improve the model, such as including additional attributes in the training data or using a different learning algorithm. Such a visualization can be used for any learning task that uses spatial data and regression learning.

6. Discovery of Quantitative Errors in the Data

Having successfully trained Cubist using data for one year, we set out to see how well an NDVI model trained on one year's data would predict NDVI for another year. We thought this exercise would serve two purposes. If we generally found transfers across years, that would be good news for Earth scientists, because it would let them use the model to obtain reasonably accurate NDVI values for years in which satellite-based measurements of NDVI are not available. On the other hand, if the model learned from one year's data transferred well to some years but not others, that would indicate some change in the world's ecosystem across those years. Such a finding could lead to clues about temporal phenomena in Earth science such as El Niños or global warming.

What we found, to our surprise, is that the model trained on 1983 data worked very well when tested on the 1984 data, and that the model trained on 1985 data worked very well on data from 1986, 1987, and 1988, but that the model trained on 1984 data performed poorly when tested on 1985 data. Table 3 shows the cross-validated correlation coefficients for each year, as well as the correlation coefficients obtained when testing each year's model on the next year's data. Clearly, something changed between 1984 and 1985. At first we thought this change might have been caused by the El Niño that occurred during that period.

Table 3. Correlation coefficients obtained when cross-validating using one year's data and when training on one year's data and testing on the next year's data, using the original data set.

DATA SET	r
CROSS-VALIDATE 1983	0.97
CROSS-VALIDATE 1984	0.97
CROSS-VALIDATE 1985	0.92
CROSS-VALIDATE 1986	0.92
CROSS-VALIDATE 1987	0.91
CROSS-VALIDATE 1988	0.91
TRAIN 1983, TEST 1984	0.97
TRAIN 1984, TEST 1985	0.80
TRAIN 1985, TEST 1986	0.91
TRAIN 1986, TEST 1987	0.91
TRAIN 1987, TEST 1988	0.90

Table 4. Correlation coefficients obtained when cross-validating using one year's data and when training on one year's data and testing on the next year's data, using the corrected data set.

DATA SET	r
CROSS-VALIDATE 1983	0.91
CROSS-VALIDATE 1984	0.91
CROSS-VALIDATE 1985	0.92
CROSS-VALIDATE 1986	0.92
CROSS-VALIDATE 1987	0.91
CROSS-VALIDATE 1988	0.91
TRAIN 1983, TEST 1984	0.91
TRAIN 1984, TEST 1985	0.91
TRAIN 1985, TEST 1986	0.91
TRAIN 1986, TEST 1987	0.91
TRAIN 1987, TEST 1988	0.90

Further light was cast on the nature of the change by examining the scatter plots that Cubist produces. In Figure 4, the graph on the left plots predicted NDVI against actual NDVI for the 1985 cross-validation run. The points are clustered around the $x = y$ line, indicating a good fit. The graph on the right plots predicted against actual NDVI when using 1985 data to test the model learned from 1984 data. In this graph, the points are again clearly clustered around a line, but one that has been shifted away from the $x = y$ equation. This shift is so sudden and dramatic that Potter's team believed that it could not have been caused by a natural phenomenon, but rather that it must be due to problems with the data.

Further investigation revealed that there was in fact an error in the data. In the data set given to to us, a recalibration that should have been applied to the 1983 and 1984 data had not been done. We obtained a corrected data set and repeated each of the Cubist runs from Table 3, obtaining the results in Table 4.[4] With the corrected data set, the model from any one year transfers very well to the other years, so these models should be useful to Earth scientists in order to provide NDVI values for years in which no satellite-based measurements of NDVI are available.

Our experience in finding this error in the data suggests a general method of searching for calibration errors in time-series data, even when no model of the data is available. This method involves learning a model from the data for each time step and then testing this model on data from successive time steps. If

there exist situations in which the model fits the data unusually poorly, then those are good places to look for calibration errors in the data. Of course, when such situations are found, the human experts must examine the relevant data to determine, based on their domain knowledge, whether the sudden change in the model results from an error in the data, from a known discontinuity in the natural system being modeled, or from a genuinely new scientific discovery. This idea can be extended beyond time-series problems to any data set that can be naturally divided into distinct sets, including spatial data.

7. Related Work

Robust algorithms for flexible regression have been available for some time. Breiman, Friedman, Olshen, and Stone's (1984) CART first introduced the notion of inducing regression trees to predict numeric attributes, whereas Weiss and Indurkhya (1993) extended the idea to rule induction. Each approach has proved successful in many domains, and both CART and Cubist have achieved commercial success. However, neither approach has yet seen much application to Earth science data, despite the considerable work on classification learning for tasks like assigning ground cover types to pixels (e.g., Brodley & Friedl, 1999) and clustering adjacent pixels into groups (e.g., Ester, Kriegel, Sander, & Xu, 1996).

The work on communicability and understandability described in this paper builds on previous work in comprehensibility. Our requirement for communicability is similar to Michalski's (1983) "comprehensibility postulate" which states that the results of computer induc-
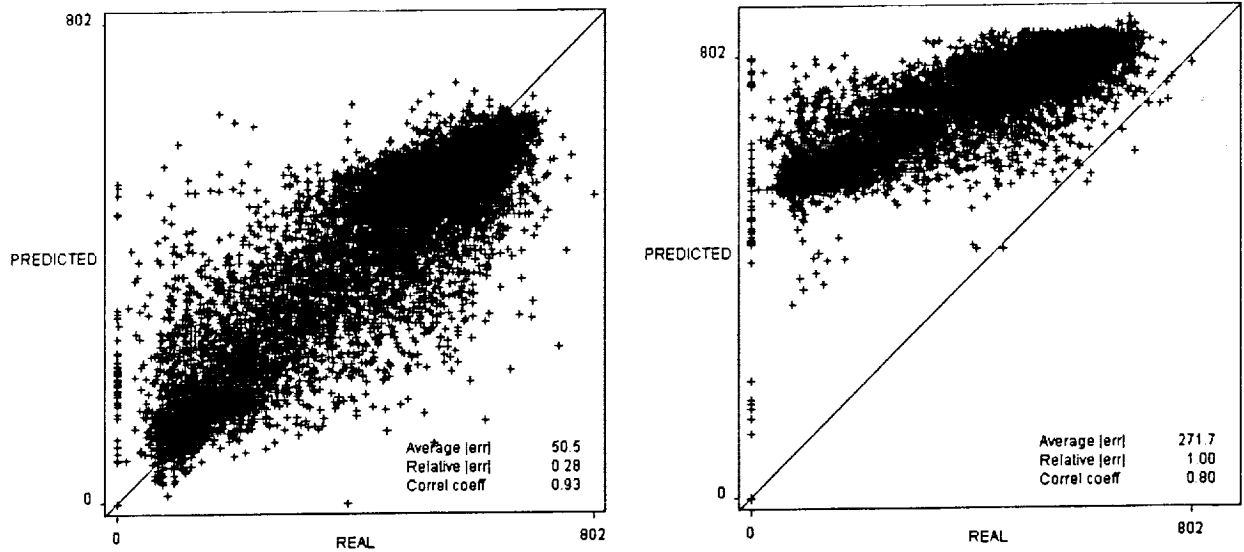
---

[4]All of the results presented in the previous sections are based on the corrected data set.

Figure 4. Predicted NDVI against actual NDVI for *(left)* cross-validated 1985 data and *(right)* training on 1984 data and testing on 1985 data.

tion should be in a form that is syntactically and semantically similar to that used by humans experts. A collection of papers on comprehensibility can be found in Kodratoff and Nédellec (1995).

Researchers have also carried out extensive work on techniques for visualizing data and learned knowledge. Tufte (1983) did early influential work on the former topic, whereas Keim and Kriegel (1996) review many of the existing approaches. Within the data-mining community, researchers have developed a variety of methods for the graphical display of learned knowledge (e.g., Brunk, Kelly, & Kohavi, 1996). However, although much of this work employs a spatial metaphor, little has focused on learned spatial knowledge itself.

Applications of machine learning to Earth science data, as in methods for ground cover prediction (e.g., Brodley & Friedl, 1999), regularly display classes on maps. Smyth, Ghil, and Ide (1999) plot predictions of a learned mixture model on the globe, but our approach to visualizing areas in which regression rules match, as well as anomalous regions, appears novel.

The European project SPIN! (2001) is seeking to develop a spatial data mining system by combining data mining tools like C4.5 (Quinlan, 1993) with tools for visualizing spatial data like Descartes (Andrienko & Andrienko, 1999). The planned system will let its users visualize geographically-referenced data on maps, and to mine the data using the data-mining

tools, from a unified user interface. The researchers plan to test the SPIN! system on applications involving seismic and volcano data. The visualization component of the project seems focused on letting users visualize the data, rather than visualizing the knowledge learned through data mining.

There has also been considerable research on using machine-learned knowledge to detect and either ignore or correct errors in training data. Much of this work has focused on removing cases with faulty class labels (e.g., John, 1995; Brodley & Friedl, 1999), but some has addressed detecting errors in the values of predictive variables. Naturally, there are established methods for detecting and correcting calibration problems in remote-sensing systems (e.g., Chen, 1997), but these rely on predefined models. Thus, our use of regression rules to detect systematic errors appears novel to both the machine learning and calibration communities.

8. Future Work

Our collaboration with Earth scientists is in its early stages, and we still have many research avenues to explore. Our next step in modeling NDVI will incorporate time explicitly by adding the year to the continuous variables used in regression equations, rather than building a separate model for each year. We hope that by examining the resulting multi-year models, we can learn something about climate change over time.

In this paper, we have assumed that models with fewer rules are more understandable. In future work, we plan to test this assumption by having our Earth science collaborators examine various sets of rules that Cubist produces for different parameter values and telling us which sets they think are easier to understand. Naturally, we will also ask them to judge the rules' plausibility and interestingness from the perspective of Earth science.

At the Potter team's suggestion, our runs with Cubist have included additional variables beyond those used in their 1998 article. Preliminary results indicate that some of these variables give small improvements in the predicted accuracy for NDVI. We plan to further investigate the utility of these variables and investigate ways to measure which variables are most important in a set of regression rules.

The NDVI predictive model is only one piece of a larger framework, known as CASA (Potter & Klooster, 1998), that Potter's team has developed to model the Earth's ecosystem. CASA takes the form of a process model, stated in terms of differential equations, for the production and absorption of biogenic trace gases in the Earth's atmosphere. For the reasons of understandability and communicability described earlier, we would like our learned models to take the same form, which means we cannot rely on Cubist alone in our future efforts.

There has been some research on discovering laws that take the form of differential equations (Todorovski & Dzeroski, 1997), but this work has not used an existing set of equations as the starting point. We plan to develop an algorithm that will begin with the current CASA model and search through the space of possible equations to find an improved model. We hope that this effort will improve the accuracy of the CASA equations while retaining its communicability and its scientific plausibility. We also hope that the changes our system makes to the model will suggest new insights about Earth science.

9. Lessons Learned

In their editorial on applied research in machine learning, Provost and Kohavi (1998) claimed that a good application paper will "focus research on important unsolved problems that currently restrict the practical applicability of machine learning methods." In this paper, we have identified, and provided initial solutions for, three such problems that arise in scientific applications:

Communicability. In scientific domains, it is important for the form of the learned models to match the form that is customarily used in the relevant literature, so that the learned models can be communicated to other scientists.

Understandability. In domains that involve spatial data, understanding of the models can be increased by visualizing the spatial distribution of the model's errors and visualizing the locations in which the model's components (e.g., rules) are active.

Quantitative errors. In applications that involve time-series numerical data, machine learning methods can be used to identify quantitative errors by testing a learned model for one time period against data from other time periods.

Although we have developed these ideas in the context of a specific scientific application – the prediction of NDVI from climate variables – we believe they have general applicability to any domain that involves scientific understanding of spatio-temporal data. As we continue utilizing machine learning to improve the CASA model, we expect that the challenging nature of the task will reveal other methods and principles that contribute to both Earth science and the science of machine learning.

Acknowledgements

References

Andrienko, G. L., & Andrienko, N. V. (1999). Interactive maps for visual data exploration. *International Journal Geographic Information Science*, *13*, 355–374.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research, 11*, 131–167.

Brunk, C., Kelly, J., & Kohavi, R. (1996). MineSet: An integrated system for data mining. *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining* (pp. 135–138). Portland: AAAI Press.

Chen, H. S. (1997). *Remote sensing calibration systems: An introduction*. Hampton, VA: A. Deepak Publishing.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining* (pp. 226–231). Portland: AAAI Press.

John, G. A. (1995). Robust decision trees: Removing outliers from data. *Proceedings of the First International Conference of Knowledge Discovery and Data Mining* (pp. 174–179). Montreal: AAAI Press.

Keim, D. A., & Kriegel, H.-P. (1996). Visualization techniques for mining large databases: A comparison. *Transactions on Knowledge and Data Engineering, 8*, 923–938.

Kodratoff, Y. & Nédellec, C. (Eds.) (1995). *Working Notes of the IJCAI-95 Workshop on Machine Learning and Comprehensibility*. Montreal, Canada.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence, 20*, 111–161.

Potter, C. S., & Brooks, V. (1998). Global analysis of empirical relations between annual climate and seasonality of NDVI. *International Journal of Remote Sensing, 19*, 2921–2948.

Potter, C. S., & Klooster, S. A. (1998). Interannual variability in soil trace gas (CO_2, N_2O, NO) fluxes and analysis of controllers on regional to global scales. *Global Biochemical Cycles, 12*, 621–635.

Provost, F., & Kohavi, R. (1998). On applied research in machine learning. *Machine Learning, 30*, 127–132.

Quinlan, J. R. (1992). Learning with continuous classes. *Proceedings of the Australian Joint Conference on Artificial Intelligence* (pp. 343-348). Singapore: World Scientific.

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

Rulequest (2001). Rulequest Research data mining tools. http://www.rulequest.com.

Smyth, P., Ghil, M., & Ide, K. (1999). Multiple regimes in Northern hemisphere height fields via mixture model clustering. *Journal of the Atmospheric Sciences, 56*.

SPIN! (2001). Spatial mining for data of public interest. http://www.ccg.leeds.ac.uk/spin.

Todorovski, L., & Dzeroski, S. (1997). Declarative bias in equation discovery. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 376–384). San Francisco: Morgan Kaufmann.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Weiss, S., & Indurkhya, N. (1993). Rule-based regression. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1072–1078). Chambery, France.