Comprehensible Knowledge Discovery: Gaining Insight from Data

Michael J. Pazzani (pazzani@ics.uci.edu)
Department of Information and Computer Science
The University of California
Irvine, CA 92697
(714) 824-5888

1. INTRODUCTION: COMPREHENSIBILITY AND DATA MINING

Large databases are routinely being collected in government, science, business and medicine. A variety of techniques from statistics, signal processing, pattern recognition, machine learning, and neural networks have been proposed to help people understand the data by discovering predictive models of the data. However, research in data mining has not paid attention to the factors that make learned models comprehensible. A comprehensible model would provide insight to an expert. This insight could be communicated to others and could support a variety of decision-making tasks. For example, knowledge acquired through such methods on a medical database might be published in scientific journals and provide advice on how to reduce the likelihood of getting a certain disease. Analysis of a political database might reveal the conditions under which trade negotiations are likely to be successful and provide advice on strategies that are effective. Knowledge acquired from analyzing a financial database might be taught in a management school.

In this paper, we argue that existing data mining systems fail to realize the full potential benefits of data mining because they do not seriously address the issue of the comprehensibility of learned models. We argue that models that are minor variants of what is already known are preferable to models that differ drastically from current understanding when the predictive power of such models are equivalent. We describe extensions to a learning system, FOCL and demonstrate that these extensions produce learned models that are more useful to experts. To illustrate our point, consider the following simple examples of economic sanctions incidents:

> ***In 1983, Australia refused to sell uranium to France, unless France ceased nuclear testing in the South Pacific. France paid a higher price to buy uranium from South Africa. A total of 1500 tons were delivered.***
>
> ***In 1980, the US refused to sell grain to the Soviet Union unless the Soviet Union withdrew troops from Afghanistan. The Soviet Union paid a higher price to buy grain from Argentina and did not withdraw from Afghanistan.***

There are a variety of lessons that one could learn from these two examples. For example, the following two "rules" are consistent with the above two examples:

If an English speaking democracy that imports oil threatens a country in the Northern Hemisphere that has a strong economic health and exports weapons, then the sanction will fail because a country in the Southern Hemisphere will sell the product.

If a country that exports a commonly available commodity tries to coerce a country with strong economic health that imports the commodity by refusing to sell them the commodity, the sanction will fail because the country will buy the commodity at a higher price from another supplier.

The above rules are of approximately the same complexity, yet the second rule makes more sense to most people. It's not the case that a person cannot literally understand the first rule. Rather, most people have difficulty understanding why such a counterintuitive pattern was identified. Most data mining systems cannot be given knowledge of economic and politics and are forced to deal entirely with the similarities and differences among examples. As a consequence, they are susceptible to producing rules whose content is incomprehensible.

Most data mining systems do not seriously address the comprehensibility of the content of learned models. Many learning methods (e.g., neural networks and logistic regression) provide accurate predictive models but provide little insight. For example, a neural network might be able to identify a fraudulent tax return, but would provide little guidance on how to change reporting requirements to reduce fraud. Symbolic learning systems (e.g., rules (Clark & Niblett, 1989, Cohen, 1995; Quinlan, 1990 and trees Quinlan, 1993)) are designed to make decision criteria explicit. However, they ignore the existing knowledge of a domain and often find decision criteria that bewilder experts. Section 3 provides more detail on this point. Most work on improving the comprehensibility of trees and rules focuses on producing concise descriptions (e.g., Karalic, 1997; Craven, 1996; Quinlan, 1987). However, this work provides no guidance on selecting among several models of the data of similar complexity. Some work on increasing the understandability of learned models concerns the construction of tools for visualizing or interactively exploring the results of learning (e.g., The MineSet Tree Visualizer- Kohavi, Sommerfield, & Dougherty, 1996). While these tools provide an excellent means of identifying and exploring what was learned, they do not provide insight unless the underlying learned models make sense to experts.

In the remainder of this paper, we show how existing knowledge may be provided to learning systems so that they are predisposed to produce models that are related to this existing knowledge. We discuss a variety of forms of background knowledge and show how this knowledge may be used to improve the comprehensibility of learned rules. We illustrate with examples from analyzing a database on foreign trade negotiations, a database on trouble reports in a telephone network, and a database on dementia provided the Consortium to Establish a Registry for Alzheimer's Disease.

2. BACKGROUND: RULE LEARNING IN FOCL

FOCL's rule learning system is derived from that of FOIL (Quinlan, 1990). FOIL learns classification rules inductively by constructing a set of rules (Horn Clauses) in terms of

predicates used to describe examples. Each clause body consists of a conjunction of predicates that cover some positive and no negative examples. FOIL starts to learn a clause body by finding the predicate with the maximum information gain, and continues to add predicates to the clause body until the clause does not cover any negative examples. After learning each clause, FOIL removes from further consideration the positive examples covered by that clause. The learning process ends when all positive examples have been covered by some clause.

FOCL (Pazzani & Kibler, 1991) extends FOIL by incorporating a compatible explanation-based learning component (Mitchell, et al., 1986). This allows FOCL to take advantage of existing knowledge provided by experts. When constructing a clause body, there are two ways that FOCL can add predicates. First, it can add predicates via the same inductive method used by FOIL. Second, it can add predicates by deriving them from the rule base. FOCL uses FOIL's information-based evaluation function to derive the most informative predicates from the rule-base and to determine whether an inductively learned predicate is more informative than one learned via explanation-based learning.

When provided no initial rule base FOCL operates exactly like FOIL. However, when provided an initial rule base, FOCL looks for patterns in the data that correct errors in the initial guidelines. In addition, the revised rules tend to be more accurate than rules formed without the help of an initial rule base. Pazzani, et al. (1991) demonstrate that FOCL can utilize incomplete and incorrect domain theories. We attribute this capability to its uniform use of an evaluation function to decide whether to include predicates learned inductively or via explanation-based learning.

We have used FOCL on a large problem from NYNEX (the parent company of New York Telephone and New England Telephone). Nynex Max (Rabinowitz, et al., 1991) is an expert system used at several sites to determine the location of a malfunction for customer-reported telephone troubles. It can be viewed as solving a classification problem where the inputs are data such as the type of switching equipment, various voltages and resistances, and the output is the location to which a repairman should be dispatched (e.g., the problem is in the customer's equipment, the customer's wiring, the cable facilities, or the central office). Nynex Max requires some customization at each site in which it is installed.

We compared the effectiveness of FOCL at customizing the Nynex Max rule-base. The existing rules are taken from one site, and the training data are the desired outputs of Nynex Max at a different site. We repeated 20 runs of each algorithm on 500 randomly selected training examples and evaluated the accuracy on an independent set of 300 test examples. The Table below shows the accuracy of the FOIL and FOCL. For comparison purposes, the accuracy of the initial domain theory is also reported. The results indicate that FOCL is more accurate than FOIL (using a paired, two-tailed t-test at the .001 level). In addition, FOCL is significantly more accurate than the original hand-crafted knowledge base.

Condition	Accuracy
Initial Rule Base	.946
FOIL	.917
FOCL	.987

The original goal of including a rule base was to improve the accuracy of the results of learning. However, in this problem we found an additional benefit: experts were more willing to use the results of learning with a rule-base because the learned rules were minor variants of what was already believed by the experts. In contrast, FOIL and other purely inductive learners produced models that differed drastically from the current understanding and practice. If there were a benefit in terms of predictive accuracy in adopting such a model, it would make sense to retrain the experts to use the new model. However, the inductive learners instead offered an alternative that was no more accurate without any explanation of why or whether it was better than the rules used by the experts.

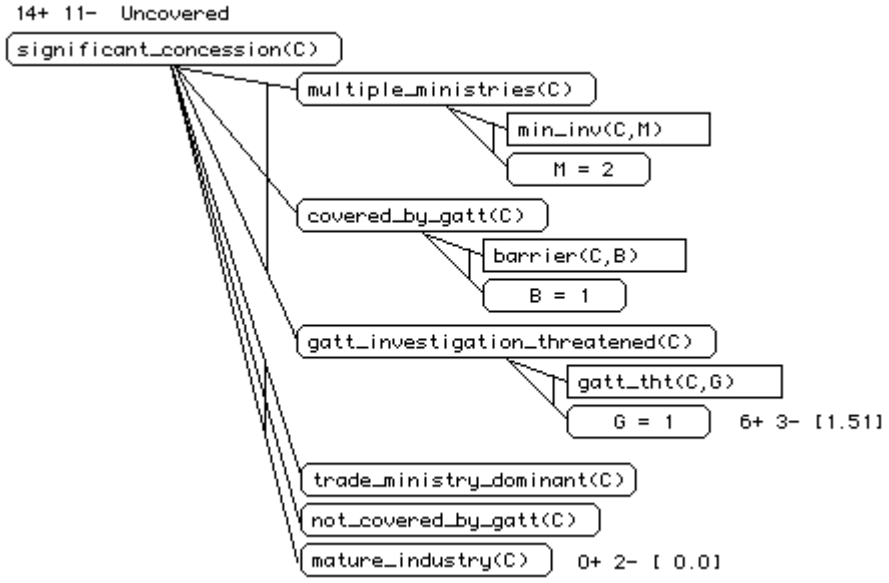
3. RULE LEARNING WITH A STRONG DOMAIN THEORY

In this study, we are dealing with a database of trade negotiations involving the United States and a trading partner. These examples serve as input to the learning programming. Each example is described by a collection of 37 variables that encode information such as the commodity being negotiated, economic data on the countries involved, and information on the two governments. Many of the variables are binary (e.g., Does the industry receive subsidies from the government); some variables take on numeric values (e.g., number of months before a major election) and a few variables are nominal (e.g., type of commodity involved). In addition, there is a binary variable associated with each case that indicates whether an agreement was reached in which the trading partner made a significant concession.

In addition to the cases, FOCL also accepts as input a rule-base that can be used to predict the classification of examples. In this application, the rules indicate the conditions under which a trading partner is predicted to make a significant concession. Three separate domain theories were developed for this problem. The first rule-base was developed by a computer science graduate student and was primarily intended to demonstrate the feasibility of using an approximate rule-base in a learning program. This rule-base correctly classified 44% of the examples. A political analyst developed the second domain theory after a review of the relevant literature. This rule-base correctly classified 61% of the examples. Finally, the political analyst produced a third rule-base by modifying the second rule-base after an analysis of the examples that were classified incorrectly by the second rule-base. The third rule-base is 71.6% accurate.

One rule indicated that there will be a significant concession in a case if the opponent has several ministries involved in the negotiation (e.g., the trade ministry and the agricultural ministry), and the issue is covered by GATT (the general agreement on tariffs and trade), and the US has threatened a GATT investigation. Another rule states that there will be a significant concession if the trade ministry is the only ministry involved in the negotiation, the issue is not covered by GATT, and the industry is mature.

FOCL has a graphical interface that allows the user to visualize which examples are correctly and incorrectly classified and summarize which rules are involved in correct and incorrect classifications. The figure below illustrates the graphical interface to FOCL on the foreign trade negotiation case. It shows that the portion of the rule base covered by GATT is the most informative (i.e., accurate) and also shows that this could be further improved.



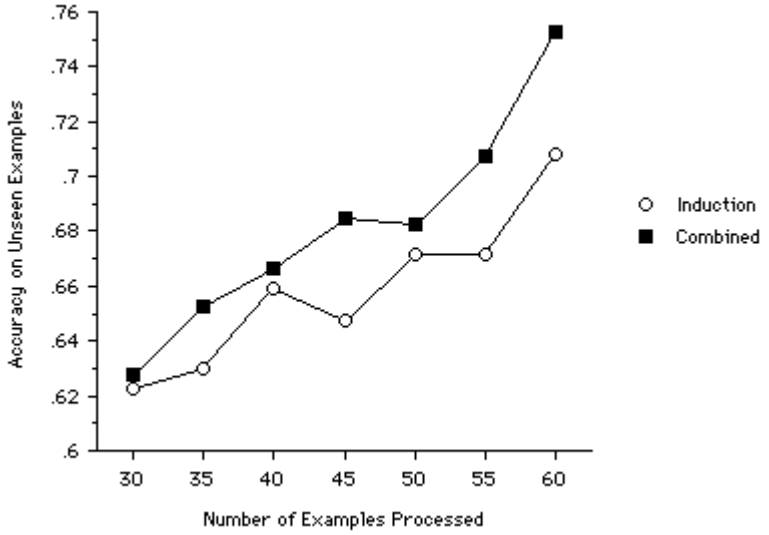
In order to access the impact of using a rule base in FOCL, the accuracy of the rules learned by FOCL was compared to the accuracy of the initial rule base. The accuracy of the learned rules is computed by a "leave-one-out". The learning algorithm is run multiple times, each time leaving out one example and recording whether the learned rule correctly classifies the example that was left out. We report the accuracy as the percentage of examples that were correctly classified by the rules learned by running FOCL on the example set with that single example removed. The accuracy of the rule base is computed by finding the percentage of examples which are correctly classified by the rules. The table below shows the result of applying the learning method under three different conditions.

1) Novice rules: These rules were created by computer science graduate students representing a nonspecialist's theory of factors that influence concessions.
2) Original expert rules: These rules were created by an experienced political scientist after a review of the literature in this area (but without examining any data).
3) Revised expert rules: These rules are the original expert rules that were revised by the political scientist after examining the training data.

Condition	**Accuracy of initial rules**	**Accuracy of learned rules**
None	NA	68.0
Novice rules	44.0	70.0
Original expert rules	61.3	73.3
Revised expert rules	72.0	81.3

The result of the combined learning system is never less accurate than the expert domain theories. In most cases, the accuracy of the combined system is substantially higher than that of the inductive system. This indicates that FOCL can improve upon expert rules by using the information gain heuristic to benefit from accurate parts of the domain theory while avoiding inaccurate expert rules.

In order to get a greater understanding of the effect of the domain theory on the learning process, we ran a simulation in which we compared the accuracy of FOCL with no domain theory to that of FOCL with the revised expert domain theory. In this simulation, we varied the size of the training set by increments of 5 from 30 to 60 examples, and computed the accuracy of the learned rules on the remaining examples. We ran 128 simulations with different randomly selected training sets of each size. The figure below plots the mean accuracy of FOCL both with and without a domain theory as a function of the size of the training set.



The figure demonstrates that the predictive accuracy of FOCL is improved when it is given an initial domain theory. This occurs because FOCL identifies the parts of the domain theory that are correct, and deletes or adds conditions as needed. There is an additional advantage to this strategy: the rules learned are more likely to be meaningful to an expert since they incorporate many of the same conditions identified by the expert.

4. RULE LEARNING WITH A WEAK DOMAIN THEORY

The second problem discussed is the analysis of a database collected by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). An electronic patient database was collected containing data on the dementia status of each patient and the results of two commonly used cognitive tests for dementia screening, the Blessed Orientation, Memory and Concentration test (BOMC- Fillenbaum et al., 1987) and the Mini-Mental Status Exam (MMSE- Folstein et al., 1975). The particular problem of interest is to identify patients with early signs of dementia.

Most demented patients do not see a physician for the problem of memory loss until four years after symptom onset. In previous research (Shankle, Mani, Pazzani, & Smyth, 1997), we have shown that a variety of machine learning and statistical methods can acquire models that have accuracy, specificity and sensitivity that exceed the average practitioner at screening for early stages of dementia. However, it is unlikely that the description of patients with early dementia created by any of the models so far would be widely adopted in the practice. The decision procedure implied by some models (e.g., logistic regression) is too complex to follow, while the decision criteria explicitly stated in learned rules or decision trees make little sense to the neurologist or the practitioner since it differs drastically from the current practice.

To understand why the results of current knowledge-discovery algorithms make little sense, it is necessary to describe how the tests are currently used for screening. In each test, the patient answers questions that assess orientation for time and place, registration, attention, short-term recall, language skills, and drawing ability. For example, the patient is first asked to remember a name and address ("John Brown, 42 Market Street, Chicago") and later asked to recall these items. The patient receives a score for each item in the test. For example, the number of times that the test giver repeats the name and address before the patient is able to repeat it immediately is recorded.

However, neither the trees produced by C4.5 (Quinlan, 1993) nor the rules produced by rule learners such as C4.5 rules or FOCL produced rules that would be acceptable in practice. In particular, some items that should be viewed as signs of being impaired are used as signs of being normal and vice versa. This does not match the original intent of the BOMC and MMSE tests, and reduces the comprehensibility of the rules to the layperson and the trained neurologist. We show one such rule below, underlining those conditions that violate the intended use of tests.

```
IF the years of education of the patient is > 5
AND the patient does not know the date
AND the patient does not know the name of a nearby street
THEN The patient is NORMAL

OTHERWISE IF the number of repetitions before correctly reciting the address is > 2
AND the age of the patient is > 86
THEN The patient is NORMAL

OTHERWISE IF the years of education of the patient is > 9
AND the mistakes recalling the address is < 2
THEN The patient is NORMAL

OTHERWISE The patient is IMPAIRED
```

If such violations of expectations were necessary to obtain accurate results, they could be tolerated. Such violations might even lead to new insights by focusing future research on explaining them. However, we shall show that on this problem, the same diagnostic performance can be achieved without these violations. We considered working with a neurologist to develop a rule-base to bias this learning task. While the expert was good at identifying which tests would be associated with dementia, the expert could not produce a rule-base consisting necessary and sufficient definition. That is, the expert could identify relevant factors and their direction, but could not say how these factors were to be combined to form an overall decision. Therefore, we decided to modify FOCL to accept this type of knowledge.

For variables with numeric relationships, the user declares whether the variable has a known monotonic relationship with each class. A monotonic relationship is one in which increasing the value of the variable always increases or decreases the likelihood category membership. When considering tests to add to a clause, the tests that violate these relationships are removed from consideration. For example, when learning a description of the normal patients, FOCL with monotonicity constraints only checks to see if the number of errors recalling the address is less than some number. When learning clauses describing the impaired category, it only tests to see if this variable is above some threshold.

These constraints on tests may also be used on Boolean and nominal variables. In this case, the user specifies which values are possibly indicative of membership in a class. For example, a value of true for the variable "knows the date" may be used as a sign for normal, while the value false may be used as a sign for impaired.

For the CERAD data, and for many medical data sets, the data is coded such that an increase in a variable's value or an incorrect response to a question increases the chance that one has a particular disease or syndrome. We encoded this knowledge as monotonicity relationships to FOCL. We also added constraints indicating that the likelihood that one is impaired increases with age and decreases with educational level. The table below shows an example of a rule learned with these constraints.

```
IF the years of education of the patient is > 5
AND the mistakes recalling the address is < 2
THEN The patient is NORMAL

OTHERWISE
IF the years of education of the patient is > 11
AND the errors made saying the months backward is < 2
THEN The patient is NORMAL

OTHERWISE
IF the years of education of the patient is > 17
THEN The patient is NORMAL

OTHERWISE The patient is IMPAIRED
```

We ran 50 trials of FOCL with and without monotonicity constraints. There is not a substantial or significant difference in accuracy in using the constraints. FOCL is 90.7% accurate when using monotonicity constraint and 90.6% accurate when unconstrained. On average, the rules formed without constraints contain a total of 4.65 tests and 2.13 violations of the monotonicity constraints.

We have conducted surveys of two neurologists to determine whether monotonicity constraints influence the willingness to follow guidelines. We generated 16 sets of rules such as that shown in Table 1 by using unconstrained FOCL and 16 rule sets such as that shown in Table 3 by using FOCL with monotonicity constraints on 16 randomly selected subsets containing 200 examples from the CERAD database. In both cases, the rule optimization procedure of FOCL was used to ensure that concise descriptions were learned. Each rule was printed on a separate sheet of paper and presented in a random order to each neurologist. We asked each neurologist to rate on a scale of 0-10 "How willing would you be to follow the decision rule in screening for

cognitively impaired patients". We hypothesized that the neurologists would be more willing to use rules that were generated by FOCL when it used monotonicity constraints.

Neurologist 1 has been involved in this project for approximately one year and is aware that the focus of the research is to create comprehensible rules. Neurologist 2 is not affiliated with this project and is unaware of its goals. For Neurologist 1, the average score of rules generated by FOCL without the monotonicity constraints was 3.25, while the average score of rules generated with the monotonicity constraints was significantly higher 5.56 $t(15) = 6.60$, $p < .001$. For Neurologist 2, these scores were 0.25 and 2.38 $t(15) = 5.09$, $p < .001$. Although it is clear that the neurologists were using different scales, in each case higher average ratings were given to the category descriptions generated with these constraints in mind. We also show the correlation between the number of monotonicity constraint violations and the willingness to follow the rule. The table below shows the correlation between these variables for each neurologist. For comparison purposes, we also show the correlation between the willingness to follow a rule and the number of tests and number of clauses in the rule, two commonly used measures of rule complexity. We did not attempt to balance for the size of the rules in the two conditions, but the average number of tests and clauses was within 10% between the two conditions.

Correlation	**Neurologist 1**	**Neurologist 2**
Violations	.433	.623
Number of tests	.208	.020
Number of clauses	.278	.011

These results show that both neurologists were sensitive to the violations of monotonicity constraints and these violations affect the willingness to follow the rule. The size of the rules did affect the judgment of one of the neurologists but to a lesser extent than the number of constraint violations.

5. CONCLUSION

We assert that the comprehensibility of learned models is an area that receives too little attention in data mining systems. Most systems equate comprehensibility with conciseness and do not acknowledge that two alternatives with the same complexity can differ in comprehensibility. We have shown that FOCL can achieve more comprehensible results when the output is biased by existing knowledge. We have demonstrated how strong domain knowledge may be represented as a rule-base and how weak domain knowledge can be represented as monotonicity constraints. We have shown that when learners are biased in this manner, the rules are more comprehensible to experts.

6. REFERENCES

Clark, P. & Matwin, S. (1993). Using Qualitative Models to Guide Inductive Learning. The *Proceedings of the 10th International Conference on Machine Learning.* Amherst, MA, 49-56.

Clark, P. & Niblett, T. (1989). The CN2 Induction Algorithm. *Machine Learning, 3,* 261-284.

Craven, M. W. (1996). Extracting Comprehensible Models from Trained Neural Networks. Ph.D. thesis, Department of Computer Sciences, University of Wisconsin-Madison.

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

Duda, R. & Hart, P. (1973). *Pattern classification and scene analysis.* New York: John Wiley & Sons.

Karalic, A. (1996). Producing More Comprehensible Models While Retaining Their Performance. Information, Statistics and Induction in Science, Melbourne, Australia.

Kohavi, R., Sommerfield D., & Dougherty J., (1996). Data Mining using MLC++, a Machine Learning Library in C++. IEEE Tools With Artificial Intelligence.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based learning: A unifying view. *Machine Learning, 1.*

Pazzani, M. (1991). The influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory & Cognition, 17, 3,* 416-32.

Pazzani, M. & Kibler, D. (1992). The utility of knowledge in inductive learning, (9):57-94.

Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies, 27,* 221-234.

Quinlan, J.R. (1990). Learning logical definitions from relations. *Machine Learning, 5,* 239–266.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, California.

Shankle, W.R., Mani, S., Pazzani, M., & Smyth, P. (1997). Detecting very early stages of dementia from normal aging with machine learning methods. *Proceedings of the 6th Conference on Artificial Intelligence in Medicine,* Europe.

BIOGRAPHY

Michael J. Pazzani, Department of Information and Computer Science, 444 Computer Science Bldg., University of California, Irvine, CA 92697-3425. Office Phone: (714) 824-5888; Office Fax: (714) 824-3976; Email address: pazzani@ics.uci.edu.

Michael J. Pazzani is a professor and department chair in Information and Computer Science at the University of California, Irvine. UC Irvine has maintained a repository of test databases used by the research community. Dr. Pazzani has been active in Machine Learning research for the past decade with numerous publications in KDD, IJCAI, AAAI, and the International Machine Learning Conference. He has served on the editorial board of Machine Learning and the Journal of Artificial Intelligence Research. Dr. Pazzani is the moderator of the machine learning list, mailed monthly to over 1000 academic and corporate researchers.