Acceptance by medical experts of rules generated by machine learning.

M. J. Pazzani[1], S. Mani[1], W. R. Shankle[2]

[1]Department of Information and Computer Science, University of California, Irvine, CA

(949) 824-7405, (949) 824-3976 (fax), pazzani@ics.uci.edu

[2]Department of Neurology, University of California, Irvine, CA

Abstract

1. Objectives:

The aim was to evaluate the potential for monotonicity constraints to bias machine learning systems to learn rules that were both accurate and meaningful.

2. Methods:

Two data sets from problems as diverse as screening for dementia and assessing the risk of mental retardation were collected and a rule learning system with and without monotonicity constraints was run on each. The rules were shown to experts who were asked how willing they would be to use such rules in practice. The accuracy of the rules was also evaluated.

3. Results:

Rules learned with monotonicity constraints were at least as accurate as rules learned without such constraints. Experts were on average more willing to use the rules learned with the monotonicity constraints.

4. Conclusions:

Analysis of medical databases has the potential of improving patient outcomes and/or lowering the cost of health care delivery. A variety of techniques from statistics, pattern recognition, machine learning, and neural networks have been proposed to “mine” this data by uncovering patterns that may be used to guide decision making. This study suggests that the cognitive factors that make learned models coherent and, therefore, credible to experts. One factor that influences the acceptance of learned models is consistency with existing medical knowledge.

Key words:

Alzheimer Disease [C10.228.140.380.100]

Mental Retardation [C23.888.592.604.646]

Artificial Intelligence - L01.700.568.110.065

Introduction

Knowledge-discovery in databases (KDD) is a field whose goal is to extract usable knowledge from a collection of data. Some recent applications in the medical domain include differential diagnosis of abdominal pain (Ohmann, 1995) and learning from a database of sports injuries (Zelic et al., 1997). In KDD, learned models are expected to be accurate and are further expected to be intelligible to experts in the field. Knowledge acquired through such methods on a medical database can be used as a hypothesis for further study and eventual publication in scientific journals or written down as a guideline to be followed in a health maintenance organization. While it is important that such knowledge is an accurate summary of the data and be verified on data not seen by the knowledge-discovery system, it is equally important that the knowledge be credible to experts in the domain. In this paper, we address the following issues:

- We argue that one factor that influences the acceptability of learned knowledge is consistency with existing medical knowledge and show that most existing knowledge-discovery algorithms too easily violate such knowledge, resulting in concepts that are not coherent when taken in the larger context of other related problems.
- We propose an algorithm that takes this prior knowledge into account. Although we implement this algorithm as an extension to the rule learner FOCL (Pazzani & Kibler, 1992), the general technique could be incorporated into other rule learners.
- We show that experts are more willing to use rules that are discovered by such a system.

Medical Databases

We consider a database collected by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). The particular problem of interest in our investigation 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 (Ernst and Hay, 1994). An electronic patient database was collected containing data on the dementia status of 305 patients[1] and the results of a commonly used cognitive test for dementia screening, the Mini-Mental Status Exam (MMSE- Folstein et al., 1975). The MMSE test is used in practice for screening for dementia by the use of a simple threshold on the total number of errors made, with the threshold depending upon the patient's age and years of education. The score on

[1] In this study, we examine only the subset of the patients that are normal (CDRS = 0) and mildy impaired (CDRS = 0.5). We eliminate those with CDRS score of 1, 2, and 3 because the more severe cases are much easier to detect in screening.

each question of the test and demographic information will be used to predict whether a patient is "normal" or "mildly impaired" by KDD methods.

We have selected the domain of Mental Retardation (MR) for our second study due to its complex nature and lack of simple and insightful models. MR has a complex etiology with an interplay of genetic and environmental factors, but the causal mechanisms are not clearly understood. Mild Mental Retardation (MMR), with an IQ range of 50-69, constitutes eighty percent of all MR and has no known biological cause in more than half of the cases (Batshaw 1993). In contrast, Severe Mental Retardation (with an IQ < 50), has a known organic basis (e.g. Down's Syndrome, anencephaly, cretinism etc.) in most instances. In this study, we have focused on learning predictive models of MMR. The database used for this study is from the National Collaborative Perinatal, of the National Institute of Neurological and Communicative Disorders and Stroke. We identified twenty variables (prenatal, perinatal and postnatal) which are thought to play a role in MR. Since our goal was generating models for early detection and intervention, we included children in the IQ range 70-84 (> 2 SD and < 1 SD) also. This category was previously referred to as borderline MR but dropped subsequently to restrict eligibility of services (e.g. special schooling) to children with IQ below 70. We included this group in the MMR group to extend MMR into a region of milder impairment resulting in 2138 cases. An equal number of controls (IQ > 85) were also randomly selected for this study.

Rule Learners

There are a variety of approaches to knowledge discovery in databases that could be applied to the above databases. Due to space limitations, we restrict our attention to rule learners. Our experts were most comfortable with this formalism for creating screening guidelines that could be written down and followed without a computer decision support system.[2] In previous work (Shankle, et al. 1997), we have shown that the rule learners are comparable in accuracy to decision trees such as C4.5 (Quinlan, 1993) and naïve Bayesian classifier (Duda & Hart, 1973) on this data. Although we will present results on the predictive power of the learned models on these data sets, our primary focus will be on improving the acceptance of learned rules while not decreasing the accuracy.

[2] Some have reported (Koneneko, 1991) that doctors like to use probabilistic models. The preference for one representational formalism over another is a complex issue that depends upon a number of factors such as prior training and visualization of the learned models. The problem we address is this paper is how to make one particular formalism, rule-based models, more comprehensible.

In this section, we describe the rule learner FOCL(Pazzani & Kibler, 1992). We go into detail on this algorithm because these details are needed to understand the source of "incomprehensible" rules. FOCL is derived from Quinlan's (1990) FOIL system. FOIL is designed to learn a set of clauses that distinguish positive examples of a concept from negative examples. Each clause consists of a conjunction of tests. For example, in the dementia domain a test might check to see whether the patient's age is less than a certain value, or whether the patient knows the day of the week.

FOIL operates by trying to find a clause that is true of as many positive examples as possible and no (or few) negative examples.[3] It then removes the positive examples explained by that clause from consideration and finds another clause to account for other positive examples. It repeats this clause learning process until all positive examples are explained by some clause. To learn a clause, FOIL first considers all possible clauses consisting of a single test. It selects the best of these according to an information-gain heuristic. The information gain heuristic favors tests that are true of many positive examples and few negative examples. Next, FOIL specializes the clause using the same search procedure and information-based heuristic, considering how conjoining a test to the current clause would improve it by excluding many negative examples and few positives. This specialization process continues until the clause is not true of any negative examples, resulting in a single clause that is a conjunction of tests.

FOCL follows the same procedure as FOIL to learn a set of clauses. However, it has a postprocessor that creates an ordered decision list from a set of unordered clauses. In a decision list, the clauses are ordered and the first clause whose test is true is used to determine class membership. To create a decision list, FOCL learns a set of clauses for each class (such as normal and impaired). The clause learning algorithm is run once for each class, treating the examples of that class as positive examples and the examples of all other classes as negative examples. This results in a set of clauses for each class. An optimization algorithm selects an ordered subset of the original clauses. The algorithm initializes the decision list to a default clause that predicts the most frequent class. Next, it iteratively tries to improve upon the current decision list with an operator that replaces the default rule with a learned clause and a new default clause. The impact is calculated of the result of adding each remaining clause to the end of the current decision list and assigning the examples that match no clause to the most frequent class of the unmatched examples. The change that yields the

highest impact in accuracy is made and the process is repeated until no change results in an improvement. [4]

One further detail is needed to understand how FOCL arrives at a decision list using rule optimization. When adding clauses to the decision list, FOCL also has the option of choosing a prefix of a learned clause. That is, if a clause such as X&Y&Z was learned, FOCL considers using X or X&Y in addition to X&Y&Z as a clause in the decision list. This can result in shorter, more general clauses. Such a clause optimization step has been shown to significantly simplify the learned concepts and improve the accuracy of the resulting decision list (e.g., Cohen, 1995).

To adjust the sensitivity of FOCL, a user can define a cost matrix that indicates the relative cost of misclassifying an example of C_i as an example of C_j. For example, to increase the sensitivity for dementia, we can have the cost of predicting normal for an impaired patient be twice the cost of calling an impaired patient normal. FOCL uses the cost matrix only in the rule optimization phase, selecting clauses that reduce misclassification cost rather than increase accuracy (Pazzani et al., 1994).

Acceptance of Learned Models

In previous research (Shankle, et al., 1997), we showed 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 and neural nets) 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. 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 MMSE tests and does not agree with the currently used procedure of totaling the number of errors.

Table 1 shows an example of a decision list that was produced by FOCL when training on patient records from the CERAD database. Similar problems occur with other rule learners such as C4.5rules. The figure shows a decision list with three conditions that violate current

[3] FOIL and FOCL use the minimum description length principle to trade-off the complexity of a rule with the number of examples covered and excluded. This prevents learning an overly complex rule to explain just a few exceptions.

Table 1: Sample rule with questionable tests underlined.

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

medical understanding being used as evidence for classifying a patient as impaired. Note that FOCL contains pruning methods to avoid overfitting by preventing irrelevant conditions to be included in rules. Nonetheless, there are still three violations of medical knowledge that are included in this rule. We ran 50 trials of FOCL of different subsets of 200 examples of the CERAD data. On average, a decision list had 2.13 tests that did not agree with the intended use of the MMSE test.

If such violations of expectations were necessary to obtain accurate results, they could be tolerated. Such violations might even lead to new insights by focussing future research on explaining them. However, we shall show that on this problem and on assessing the risk of mental retardation, the same diagnostic performance can be achieved without these violations. We first analyze the source of the problem and next present a solution.

If we assume that the medical knowledge is correct, then there are two factors that contribute to a test that violates these constraints being used in a rule. First, while the test appeared best according to an information-based selection procedure, this procedure detected a "spurious correlation" in the data due to sampling biases, noise in category label (i.e., a patient may be misdiagnosed) or noise in a variable's value (i.e., a question may have been recorded or scored improperly or a patient may have guessed the correct answer to a question such as the

[4] When we report on the accuracy of FOCL, we always evaluate the accuracy of learned rules on a set of test examples

day of the week). Such problems are more likely to occur near the end of a clause. Second, the selection of tests that violate the monotonicity constraints is that the selection procedure selects a single best test. Often, several tests are equally informative or statistically indistinguishable.

In the remainder of this paper, we describe a simple extension to FOCL that prevents it from learning rules that violate the expectations of a domain expert and show that this extension does not hurt the diagnostic value of the concepts that are learned. We present evidence that experts prefer rules learned with this constraint in mind.

Monotonicity Constraints

As Table 1 illustrates, some clauses violate the intent of the MMSE examination. In particular, getting some questions right is used as evidence that one is impaired and getting some questions wrong is used as evidence that one is not impaired. Similar problems occurred in the mental retardation domain. A relatively simple change to FOCL eliminates such tests from learned rules by having the user (optionally) specify the intended relationship between an attribute value and a classification. 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 tends to increase or decrease the likelihood of class membership. Tests that violate these relationships are not considered, when searching for tests to add to a clause. For example, the constraint expressed as `(increase recall_error impaired)` indicates that 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 and when learning clauses describing the impaired class, it only tests to see if this variable is above some threshold. The threshold is not specified in advance by the expert. Rather, the threshold that best distinguishes positive examples from negative examples according to the information gain criteria is selected.

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.

---

that does not include the examples used for learning or ordering.

Table 2. A rule learned with monotonicity 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

The constraints used in this paper were developed by the authors, but we believe they would be obvious to anyone familiar with the problem. In more complex situations, they might require research and independent validation. 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. Table 2 shows an example of a decision list learned with these constraints on the same data that was used to learn the decision list in Table 1. In subsequent sections, we will show that rules learned with monotonicity constraints are at least as accurate as rules learned without them, and that these rules are more acceptable to medical experts.

Violations of the Monotonicity Constraints

So far, we have assumed that the monotonicity constraints are correct and the learning system does not allow violations of the constraints. Ideally, we would not allow clauses that violate the constraints unless violating them results in more accurate decision lists. Here we describe a simple extension to FOCL that implements this idea. The decision list creation algorithm selects from a pool of clauses that contains clauses learned on the training set with constraints and clauses learned from the same training data without constraints. The decision list

ordering procedure is changed to prefer clauses learned with constraints unless a clause without constraints results in a greater increase in accuracy (as measured on the set of data reserved for ordering). By relaxing the constraints in this manner, they are used as a preference bias. We use stochastic search to generate pools of clauses consistent with the monotonicity constraints and pools without this constraint. Rather then selecting the most informative condition to add to each clause, FOCL selects among the *k* (3) most informative tests with probability proportional to the informativeness of the test. By repeating the process of learning a set of rules from the training data, several alternative partitions of the data are formed. In the experiment reported below, 5 rule sets are learned without monotonicity constraints and 5 rule sets are learned with monotonicity constraints. These are all entered into the pool of clauses for decision list creation. Note that increasing the pool of clauses does not increase the complexity of the decision list learned since most of the clauses are not used in the final decision list. Rather, this provides a richer set of possibilities for the decision list creation algorithm.

Evaluation on the Two Medical Databases

On each of the two medical databases, we will report on the predictive power of models learned by FOCL without monotonicity constraints (simply called FOCL in the remainder of the paper), FOCL with monotonicity constraints (FOCL-m) and FOCL with stochastic search and a preference bias (FOCL-pref). In each domain, we will also generate decision lists using FOCL and FOCL-m and ask experts to judge the output of the learned models.

Table 3 shows the accuracy on the CERAD data. The accuracy is averaged over 50 trials of dividing the data into a training set of size 210 and a test set of size 105. The test set does not contain any examples from the training set.

Table 3. Accuracy at identifying impaired patients.

Algorithm	**Accuracy**
FOCL	90.6
FOCL-m	90.7
FOCL-pref	94.5

Several results are worthy of highlighting from Table 3. First, the monotonicity constraints do not decrease the accuracy of FOCL, showing that the average of 2.13 constraint violations

produced by unconstrained FOCL are unnecessary. A trusting user of knowledge discovery systems such as FOCL (or other KDD algorithms such as C4.5) might be tempted to begin a research program to prove that the conventional wisdom is wrong and seek an explanation for the violations of monotonicity constraints. However, since FOCL-m has the same accuracy as FOCL, there is not yet sufficient evidence to begin such a research program.

Table 3 also shows that there is an added benefit in selecting from multiple sets of clauses learned with and without monotonicity constraints. Decisions lists learned with FOCL-pref are significantly more accurate (at the .01 level using a paired two-tailed t-test) than those learned by FOCL-m. Furthermore, the average number of monotonicity constraint violations is significantly reduced to 0.75 with FOCL-pref. In FOCL-pref, a single constraint violation occurs often. The same variable is used in 32 of the 50 runs in a manner that violates the monotonicity constraints. With the evidence that FOCL-pref is more accurate with this constraint violation, and that the single violation occurs frequently, we will be conducting a further investigation of this single variable.

We have conducted surveys of two neurologists to determine whether monotonicity constraints influence the willingness to follow guidelines. We generated 8 decisions lists such as the one shown in Table 1 by using unconstrained FOCL and 8 decision lists such as that shown in Table 2 by using FOCL with monotonicity constraints on randomly selected subsets of the CERAD database containing 200 examples. Each decision list 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 did not indicate an interpretation for each point on the scale and therefore analyzed the data for each expert separately. 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 acceptable 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 at 5.56 using a one-tailed t-test, $t(15) = 6.60$, $p < .001$. For Neurologist 2, these values were 0.25 and 2.38, $t(15) = 5.09$, $p < .001$. In each case higher average ratings were given to the category descriptions generated with these constraints in mind. These results show that both neurologists were

sensitive to the violations of monotonicity constraints and these violations affect the willingness to follow the rule.

Figure 1 shows a ROC curve comparing a threshold only on the total MMSE score to decision lists learned by FOCL-m for various misclassification cost settings (averaged over 50 trials training on 210 examples and evaluated on 105 examples). When the cost of misclassifying an impaired patient is high, the sensitivity is increased at the expense of increasing the number of false positives. The fact that the learned rules achieve higher sensitivity than the total MMSE score provides evidence that different questions of the MMSE have different diagnostic values and a simple sum of errors obscures this information. Furthermore, since the learned rules reference only a subset of the questions, it might be possible to reduce the amount of time and money spent screening for dementia.

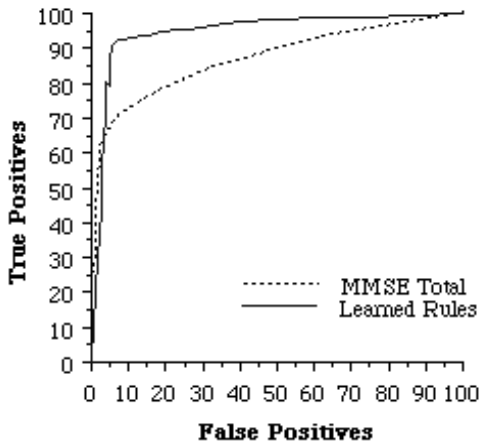

Figure 1. A ROC curve comparing a threshold on the MMSE test to decision lists learned with a variety of misclassification costs.

We ran a similar set of experiments with the database from the National Collaborative Perinatal Project. In this case, we ran 50 trials of each algorithm, training on 1000 examples and testing on 500. The average accuracy of each algorithm is reported in Table 4. All of the algorithms perform similarly, indicating again that the violations of the monotonicity constraints are not needed to achieve increased predictive power. For all of the algorithms, the learned rules were more complex in the mental retardation domain than the dementia domain. On average, there were 24.2 tests in the decision list for FOCL (6.5 clauses) and 22.0 tests for FOCL-m (also 6.5 clauses). For FOCL, there was an average of 8.3 monotonicity violations per rule. FOCL-m had similar predictive performance with no constraint

violations. FOCL-pref had an average of 0.35 constraint violations per rule showing that the additional search and the additional freedom of FOCL-pref to violate monotonicity constraints when necessary is not needed in this problem.

Table 4. Accuracy at screening for mental retardation.

Algorithm	**Accuracy**
FOCL	68.4
FOCL-m	69.2
FOCL-pref	68.5

We showed six decision lists learned by FOCL and six by FOCL-m to four experts on mental retardation. We reduced the number of rules rated by the experts in this domain because the rules were longer. None of these experts are involved in this research project or acquainted with its goals. The results of these experiments are displayed in Table 5. For two of the experts there was a significant effect on acceptability of constraint violations (using a one tailed t-test), for one expert there was a marginally significant effect, and there was no effect for another.

Table 6. Expert opinion of mental retardation rules.

	E1	**E2**	**E3**	**E4**
Acceptability (FOCL)	2.1	1.5	2.1	5.7
Acceptability (FOCL-m)	6.5	2.7	5.7	6.0
Significance level	.005	.06 MS	.01	.12 NS

Related Work

Most work in producing understandable rules has focused on syntactic properties of the rules, particularly the size of rules. Such work typically equates size with comprehensibility and seeks to minimize the size of learned relationships. For example, Karalic's (1996) paper, "Producing more comprehensible models while retaining their performance" might just as well be entitled "Producing smaller models while retaining their performance" since it describes the use of the minimum description length principle to learn shorter rules. Bohenec

and Bratko (1994) present a framework for trading off accuracy and simplicity. While we agree that unnecessarily complex models should be avoided, there are often a variety of models with similar complexity and other factors are needed to select among these alternatives. In contrast, we have focused on how the relationship between learned knowledge and existing knowledge affects acceptance and have shown that there are differences other than size that affect the willingness of experts to use rules.

Dehaspe, van Laer, and De Raedt (1994) present a general framework for constraining rule learners with a declarative language bias. It should be straightforward to implement monotonicity constraints in this bias. Our central contribution in this paper is to show that people prefer rules learned with monotonicity constraints and not on the algorithmic details of the implementation.

Research at the intersection of machine learning and knowledge acquisition, (e.g., Sleeman, & Corruble, 1998; Ganascia, Thomas & Laublet 1993; Pazzani & Brunk, 1991) has often looked at learning in the context of existing knolwedge. The focus of that line of work is often to complete missing parts of a knowledge base. In contrast, here the knowledge that constrains the learning process is more general than the knowledge acquired.

The most similar work to the research reported in this paper is Clark & Matwins'(1993) extensions to CN2 with qualitative models. We build upon this prior research by presenting a simpler formalism (monotonicity constraints) that is more appropriate for medical diagnostic domains, by optionally using this knowledge as a preference bias rather than a selection bias, and by presenting evidence that medical experts do indeed prefer rules that do not violate monotonicity constraints.

Conclusion

We have argued that conforming to monotonicity constraints makes the results of learning more acceptable to experts. By acceptable, we mean that the expert agrees that the regularity expressed in the rule could be a predictive model. KDD systems are in a sense myopic in that they work on a single problem at a time and don't have knowledge of the richer context in which they are discovering regularities. As a consequence, they can find patterns that are not coherent in a larger context. Monotonicity constraints are one simple way to express some of this larger context. Although implemented in FOCL, it should be straightforward to add to any rule learning system. In the dementia domain, these constraints represent the simple fact that persons with dementia are expected to perform worse on cognitive tests than persons without dementia. In the mental retardation domain, these constraints come from prior

knowledge of the factors that individually increase the probability of mild mental retardation. This knowledge is not complete, in the sense that it does not indicate which combinations of these factors are necessary and sufficient. Instead, they bias FOCL-m to produce rules that are consistent with the monotonicity constraints, or bias FOCL-pref to prefer rules that do not violate the constraints.

While we feel it is important for a KDD system not to violate constraints unless necessary, we also feel that violations of the existing knowledge, when supported by sufficient evidence are an important way to generate hypotheses for further study. Such violations can represent cases where the existing knowledge is incorrect. When such violations are rare and reliable, they can initiate an inquiry into explaining them. When such violations are common and unstable, they can lead to experts unwilling to use the results of KDD.

References

Batshaw, M. (1993). *Mental Retardation, volume 40 of Pediatric Clinics of North America-The Child With Developmental Disabilities*, pages 507-522. W. B. Saunders Company, Philadelphia, PA.

Bohenec, M., and Bratko, I. (1994). Trading Accuracy for Simplicity in Decision Trees. *Machine Learning, 15*, 223-250.

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.

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

Dehaspe, L., van Laer, W., and De Raedt, L. (1994). Applications of a logical discovery engine. In *Proceedings of the Fourth International Workshop on Inductive Logic Programming (ILP-94).*

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

Ernst, R. & Hay, J. (1994). The US economic and social costs of Alzheimer's disease revisited. *American Journal of Public Health,*84(8):1261-4.

Folstein, M., Folstein, S., and McHugh, P. (1975). Mini-mental state-a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3):189-98.

Ganascia, J.-G., Thomas, J., Laublet, P. (1993)- *Integrating Models of Knowledge and Machine Learning*. European Conference on Machine Learning, Wien, Austria.

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

Kononenko, I. (1991). Semi-naïve Bayesian classifier. In *Proceedings of the Sixth European Working Session on Learning*, 206-219. Berlin: Springer- Verlag.

Ohmann, C., Yang, Q., Moustakis, V., Lang, K., and Elk, V. P. (1995). Machine learning techniques applied to the diagnosis of acute abdominal pain. In Barahona, P. and Stefanelli, M., *editors, Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME95*, volume 934, pages 276--281. Springer.

Pazzani, M., & Brunk, C. (1991). Detecting and correcting errors in rule-based expert systems: an integration of empirical and explanation-based learning. *Knowledge Acquisition, 3,* 157-173.

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

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T. Brunk, C. (1994*). Reducing Misclassification Costs*. Proceedings of the 11th International Conference of Machine Learning, New Brunswick. Morgan Kaufmann, 217-225.

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.* The Proceedings of the 6th Conference on Artificial Intelligence in Medicine.

Sleeman, D. & Corruble, V. (1998). *The Role of Knowledge in a Data Mining Algorithm*. In Proceedings of the Fourth International Workshop on Multi-StrategyLearning (MSL-98). Department of Informatics, University of Torino, Vol 1, pp165-174.

Zelic, I., Kononenko, I., Lavrac, N., and Vuga, B. V. (1997). Machine learning applied to diagnosis of sport injuries. In Keravnou, E., Garbay, C., Baud, R., and Wyatt, J., editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pages 138-144. Springer.