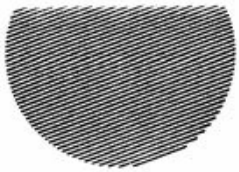

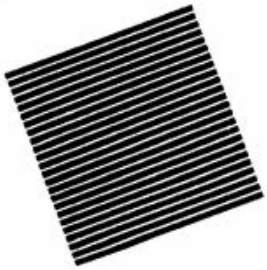

proceedings

aaai-86

fifth national conference on artificial intelligence

august 11–15, 1986 philadelphia, pa

volume 2 engineering

THE MULTI-PURPOSE INCREMENTAL LEARNING SYSTEM AQ15 AND ITS TESTING APPLICATION TO THREE MEDICAL DOMAINS*

Ryszard S. Michalski, Igor Mozetic**, Jiarong Hong***, Nada Lavrac**,

Department of Computer Science
University of Illinois at Urbana-Champaign

ABSTRACT

AQ15 is a multi-purpose inductive learning system that uses logic-based, user-oriented knowledge representation, is able to incrementally learn disjunctive concepts from noisy or overlapping examples, and can perform constructive induction (i.e., can generate new attributes in the process of learning). In an experimental application to three medical domains, the program learned decision rules that performed at the level of accuracy of human experts. A surprising and potentially significant result is the demonstration that by applying the proposed method of cover truncation and analogical matching, called TRUNC, one may drastically decrease the complexity of the knowledge base without affecting its performance accuracy.

I INTRODUCTION

It is widely acknowledged that the construction of a knowledge base represents the major bottleneck in the development of any AI system. An important method for overcoming this problem is to employ inductive learning from examples of expert decisions. In this knowledge acquisition paradigm, knowledge engineers do not have to force experts to state their "know how" in a predefined representational formalism. Experts are asked only to provide correct interpretation of existing domain data or to supply examples of their performance. It is known that experts are better at providing good examples and counterexamples of decisions than at formalizing their knowledge in the form of decision rules. Early experiments exploring this paradigm have also shown that decision rules formed by inductive learning may outperform rules provided by human experts [Michalski & Chilausky 80; Quinlan 83].

An important part of the development of an inductive learning systems is its evaluation on practical problems. There are several criteria for evaluating inductive learning methods. We argue that the most important one is the *classification accuracy* of the induced rules on new objects. In the paper we present an experimental evaluation of the AQ15 program for learning from examples in three medical domains: lymphography, prognosis of breast cancer recurrence, and location of primary tumor. These three domains are characterized by consecutively larger amounts of overlapping and sparse learning events. Examples of a few hundred patients with known diagnoses were available, along with the assessed classification accuracy of human experts. We randomly selected 70% of examples for rule learning and used the rest for rule testing. For each domain, the experiment was repeated four times. The induced rules reached the classification accuracy of human experts. Performance of experts was measured in two out of three domains, (breast cancer and primary tumor) testing four and five experts, respectively. The experiments also revealed the interesting phenomenon that by truncating covers and applying analogical rule matching one may significantly reduce the size of the knowledge base without decreasing its performance accuracy. A more detailed presentation of the results and of the program AQ15 is in [Michalski, Mozetic & Hong 86; Hong, Mozetic & Michalski 86].

II AN OVERVIEW OF AQ15

The program AQ15 is a descendant of the GEM program and the AQ1-AQ11 series of inductive learning programs, e.g., [Michalski & Larson 75]. Its ancestors were experimented with in the areas of plant disease diagnosis [Michalski & Chilausky 80, chess end-games, diagnosis of cardiac arrhythmias [Mozetic 86], and others.

All these systems are based on the AQ algorithm, which generates decision rules from a set of examples, as originally described in [Michalski 69; Michalski & McCormick 71]. When building a decision rule, AQ performs an heuristic search through a space of logical expressions to determine those that account for all positive examples and no negative examples. Because there are usually many such *complete* and *consistent* expressions [Michalski 83], the goal of AQ is to find the most preferred one, according to a flexible extra-logical criterion. This criterion is defined by the user to reflect the needs of the application domain.

Rules are represented as expressions in variable-valued logic system 1 (VL_1), which is a multiple-valued logic propositional calculus with typed variables [Michalski & Larson 75]. In VL_1, a *selector* relates a variable to a value or a disjunction of values. A conjunction of selectors forms a *complex*. A *cover* is a disjunction of complexes describing all positive examples

*This research was supported in part by the National Science Foundation under Grant No. DCR 84-06801, the Office of Naval Research under Grant No. N00014-82-K-0186, the Defense Advanced Research Project Agency under Grant No. N00014-K-85-0878, and by the Slovene Research Council.

**On leave from: *Jozef Stefan Institute, Ljubljana, Yugoslavia.*

***On leave from: *Harbin Institute of Technology, Harbin, The People's Republic of China.*

and none of the negative examples of the concept. A cover defines the condition part of a corresponding decision rule.

AQ15 is able to produce rules of different degrees of *generality* (rules may be general, minimal or specific). The program implements *incremental learning* with perfect memory. The user may supply his decision hypotheses as initial rules. In this type of learning the system remembers all learning examples that were seen so far, as well as the rules it formed [Reinke & Michalski 86]. A form of *constructive induction* is implemented in AQ15 as well. The program's background knowledge, expressed in the form of rules, is used to generate new attributes not present in input data. The background knowledge rules are of two types: L–rules that define values of new variables by logical assertions, and A–rules that introduce new variables as arithmetic functions of original variables.

III TRUNCATION OF COVERS AND ANALOGICAL MATCHING

The underlaying idea behind the TRUNC method is that the meaning of a concept can be distributed between its explicit representation and the method of its interpretation [Michalski 86a, Michalski 86b]. This idea can be simply realized as described below.

In AQ15 a concept is represented in the form of a simple conjunctive statement (called a *complex*), or as a disjunction of such statements. Each statement is associated with a pair of weights: t and u, representing the *total* number of instances (events) explained by the expression, and the number of events explained *uniquely* by that expression, respectively. The t–weight may be interpreted as a measure of the representativeness of a complex as a concept description. The u–weight may be interpreted as a measure of importance of the complex. The complex with the highest t–weight may be interpreted as describing the most typical examples of the concept. It may also be viewed as a prototypical or the ideal definition of the concept. The complexes with lowest u–weights can be viewed as describing rare, exceptional cases. If the learning events from which rules are derived are noisy, such "light" complexes may be indicative of errors in the data.

Two methods of recognizing the concept membership of an instance are distinguished: the *strict* match and the *analogical* match. In the strict match, one tests whether an instance satisfies condition part of a rule. In the analogical match, one determines the degree of similarity or conceptual closeness between the instance and the condition part. Using the strict match, one can recognize a concept without checking other candidate concepts. In the analogical match, one needs to determine the most closely related concept. The analogical matching can be accomplished in a variety of ways, ranging from approximate matching of features to *conceptual cohesiveness* [Michalski & Stepp 83].

The above weight–ordering of complexes suggests an interesting possibility. Suppose we have a t–weight ordered disjunction of complexes, and we remove from it the lightest complex. So truncated description will not strictly match events that uniquely satisfy the truncated complex. However, by applying the analogical match, these events may still come out to be the most similar to the correct concept, and thus be correctly recognized. A truncated description is of course simpler, but carries a potentially higher risk of recognition error, and requires a somewhat more sophisticated evaluation. We can proceed further and remove the next "light" complex from the cover, and observe the performance. Each such step produces a different trade–off between the complexity of the description on one side, and the risk factor and the evaluation complexity on the other (Figure 1). At some step the best overall result may be achieved for a given application domain. This method of knowledge reduction by truncating ordered covers and applying analogical matching is called TRUNC.

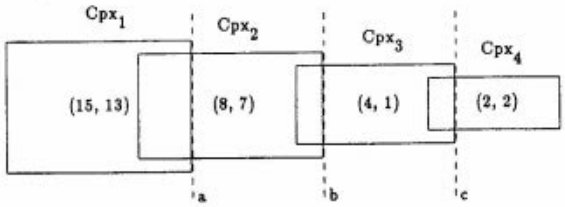

Figure 1. An example of a t–ordered cover. The cuts at a, b and c mark truncated covers with 1, 2 or 3 complexes, respectively. In each pair (x, y), x represents the t–weight, and y represents the u–weight.

The above trade–off is related to the issues studied in Variable Precision Logic [Michalski & Winston 86]. An interesting problem is to test how the cover truncation method affects the accuracy of recognition and the complexity of the decision rules in different practical settings. Section IV presents results of some such experiments, which in some cases came out very surprising. We now turn to the problem of analogical matching, and the resolution of conflict when several concept descriptions are matched by an event.

When strictly matching a new event against a set of (disjunctive) rules, three outcomes are possible: there may be only one match, more than one, or no match (categories called SINGLE, MULTIPLE and NO_MATCH, respectively; Figure 2). Each category requires a different evaluation procedure, and a different method of determining the accuracy of concept recognition. For exact match (category SINGLE), the evaluation is easy: the decision is counted as correct if it is equal to the known classification of the testing object and as wrong otherwise. If there are several exact matches (the MULTIPLE case) or none (the NO_MATCH case) the system activates the *flexible evaluation scheme* that determines the best decision (or the most probable one). Comparing this decision with the decision provided by experts, one evaluates it as correct or incorrect. Here we propose two simple heuristic classification criteria, one for the MULTIPLE case, and the other for the NO_MATCH case.

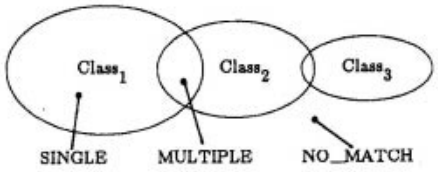

Figure 2. The three possible cases when matching a new event against a set of decision rules.

Estimate of Probability for the MULTIPLE case (EP). Let C_1, ... ,C_n denote decision classes and e an event to be classified. For each decision class C_i we have a rule that consists of a disjunction of complexes (Cpx), which, in turn are conjunctions of selectors (Sel). We define the estimate of probability, EP, as follows:

1) EP of a complex Cpx_j is the ratio of the weight of the complex (the number of learning examples covered by it) by the total number of learning examples ($\#examples$), if the complex is satisfied by the event e, and equals 0 otherwise:

$$EP(Cpx_j,e) = \begin{cases} Weight(Cpx_j) \,/\, \#examples & \text{if complex } Cpx_j \text{ is satisfied by e,} \\ 0 & \text{otherwise.} \end{cases}$$

2) EP of a class C_i is the probabilistic sum of EPs of its complexes. If the rule for C_i consists of a disjunction of two complexes $Cpx_1 \vee Cpx_2$, we have:

$$EP(C_i,e) = EP(Cpx_1,e) + EP(Cpx_2,e) - EP(Cpx_1,e)\,EP(Cpx_2,e)$$

The most probable class is the one with the largest EP, i.e., the one whose satisfied complexes cover the largest number of learning examples. Obviously, if the class is not satisfied by the given event, its EP equals 0.

Measure of Fit for the NO_MATCH case (MF). In this case the event belongs to a part of the decision space that is not covered by any decision rule and this calls for analogical matching. One way to perform such matching is to measure the fit between attribute values in the event and the class description, taking into consideration the prior probability of the class. We used in the experiments a simple measure, called *measure of fit*, MF, defined as follows:

1) MF of a selector Sel_k is 1, if the selector is satisfied. Otherwise, this measure is proportional to the amount of the decision space covered by the selector:

$$MF(Sel_k,e) = \begin{cases} 1 & \text{if selector } Sel_k \text{ is satisfied by e,} \\ \dfrac{\#Values}{DomainSize} & \text{otherwise.} \end{cases}$$

where $\#Values$ is the number of disjunctively linked attribute values in the selector, and $DomainSize$ is the total number of the attribute's possible values.

2) MF of a complex Cpx_j is defined as the product of MFs for a conjunction of its constituent selectors, weighted by the proportion of learning examples covered by the complex:

$$MF(Cpx_j,e) = \prod_k MF(Sel_k,e) \times (Weight(Cpx_j) \,/\, \#examples)$$

3) MF of a class C_i is obtained as a probabilistic sum for a disjunction of complexes.

$$MF(C_i,e) = MF(Cpx_1,e) + MF(Cpx_2,e) - MF(Cpx_1,e)\,MF(Cpx_2,e)$$

We can interpret the measure of best fit of a class as a combination of "closeness" of the event to a class and an estimate of the prior probability of the class. This measure can be further extended by introducing a measure of degree to which a selector is satisfied [Michalski & Chilausky 80].

IV EXPERIMENTS AND ANALYSIS OF RESULTS

The experiments were performed on data from three medical domains: lymphography, prognosis of breast cancer recurrence and location of primary tumor (Table 1). All data were obtained from the Institute of Oncology of the University Medical Center in Ljubljana, Yugoslavia [Kononenko, Bratko & Roskar 86].

Lymphography. This domain is characterized by 4 decision classes (diagnoses) and 18 attributes. Data of 148 patients were available. Diagnoses in this domain were not verified and actual testing of physicians was not done. A specialist's estimation is that internists diagnose correctly in about 60% and specialists in about 85% of cases.

Prognosis of Breast Cancer Recurrence. The domain is characterized by 2 decision classes and 9 attributes. The set of attributes is *incomplete* as it is not sufficient to completely discriminate between cases with different outcome. Data for 286 patients with known diagnostic status 5 years after the operation were available. Five specialists that were tested gave a correct prognosis in 64% of cases.

Location of Primary tumor. Physicians distinguish between 22 possible locations of primary tumor. Patients' diagnostic data involve 17 attributes (this set is also *incomplete*). Data of 339 patients with known locations of primary tumor were available for the experiment. Four internists that were tested determined a correct location of primary tumor in 32% of cases and four oncologists (specialists) in 42% of test cases.

Domain	Examples	Classes	Attrs	Vals/Attr
Lymphography	148	4	18	3.3
Breast cancer	286	2	9	5.8
Primary tumor	339	22	17	2.2

Table 1. The table presents the number of examples, of classes, of attributes, and the average number of values per attribute for each of the three medical domains.

In all medical domains 70% of examples were selected for learning and the remaining 30% for testing. Each testing experiment was repeated 4 times with randomly chosen learning examples. Final results are the average of 4 experiments (Table 2).

In addition to results obtained from using complete (untruncated) rules, results of two other experiments are presented. In the first experiment we eliminated from rules all complexes that cover uniquely only one learning example (unique >1), and in the second we eliminated all complexes except the most representative one covering the largest number of learning examples (best cpx). Complexity of rules is measured by the number of selectors and complexes.

Table 2 shows that some results came out very surprising. When the cover of each class was truncated to only one (the heaviest) complex, the complexity of the rule set for lymphography went down from the total of 12 complexes and 37 selectors to only 4 complexes (one per class) and 10 selectors (see bold numbers). At the same time the performance of rules

Domain	Cover truncation	Complexity Sel	Cpx	Accuracy	Human Experts	Random Choice
Lymphography	no	**37**	**12**	**81%**	85% (estimate)	25%
	unique >1	34	10	80%		
	best cpx	**10**	**4**	**82%**		
Breast cancer	no	**160**	**41**	**66%**	64%	50%
	unique >1	128	32	66%		
	best cpx	**7**	**2**	**68%**		
Primary tumor	no	551	104	39%	42%	5%
	unique >1	257	42	41%		
	best cpx	112	20	29%		

Table 2. Average complexity and accuracy of AQ15's rules learned from 70% of examples, over 4 experiments, as compared to the performance of human experts and a random choice classification algorithm.

went slightly up (from 81% to 82%)! A similar phenomenon occurred in the breast cancer domain, where the number of selectors and complexes went down from 160 and 41 to 7 and 2, respectively; while the performance went slightly up from 66% to 68%. This means that by using the TRUNC method one may significantly reduce the knowledge base without affecting its performance accuracy. Results for human experts are the average of testing of five and four domain specialists in the domains of breast cancer recurrence and primary tumor, respectively [Kononenko, Bratko & Roskar 86]. In the domain of lymphography, physicians' accuracy is given only as their estimate and was not actually measured.

The domain of lymphography seems to have some strong patterns and the set of attributes is known to be complete. There are four possible diagnoses but only two of them are prevailing. The domain of breast cancer has only two decision classes but does not have many strong patterns. Domain of location of primary tumor has many decision classes and mostly binary attributes. There are only a few examples per class, and the domain seems to be without any strong patterns. Both domains are underspecified in the sense that the set of available attributes is incomplete (not sufficient to discriminate between different classes). The statistics in Table 3 include average number of complexes per rule, average number of attributes per complex, average number of values per attribute and finally, average number of learning examples covered by one complex. We can see that in the domain of primary tumor decision rules consist of complexes that in average cover slightly more than 2 examples. In the domain of lymphography complexes in average cover 8 examples, which indicates a presence of relatively strong patterns.

It is surprising that a cover truncation mechanism that strongly simplifies the rule base may have no effect on classification accuracy. Removing "light" complexes from a cover is equivalent to removing disjunctively linked conditions from a concept description. This process thus overspecializes a knowledge representation, producing an *incomplete* concept description (i.e., a one that does not cover some positive examples). As the results show, this may lead to a substantial simplification of the concept description, without the decline in performance of the rules base.

This knowledge reduction technique by specialization may be contrasted with knowledge reduction by generalization used in the ASSISTANT learning program, a descendant of ID3 [Quinlan 83]. This program represents knowledge in the form of decision trees, and has been applied to the same medical problems as here [Kononenko, Bratko & Roskar 86]. The program applies a *tree pruning* technique based on the principle of maximal classification accuracy. The technique removes certain nodes from a tree, and is equivalent to removing conjunctively linked conditions from a concept description. Thus, such a knowledge reduction technique overgeneralizes the knowledge representation, producing an *inconsistent* concept description (i.e., a one that covers some negative examples). It is interesting to point out that this technique may also lead to an improvement of accuracy in decision making when learning from noisy and overlapping data. Table 4 presents the complexity and diagnostic accuracy of ASSISTANT's trees built with and without the tree pruning mechanism [Kononenko, Bratko & Roskar 86].

Tree pruning corresponds to the removal of selectors from complexes. This seems to suggest that when learning from noisy or overlapping data the knowledge reduction process may not only involve removal of complexes from a cover (a specialization process) but also removal of selectors from complexes (a generalization process). This means that a concept description would be both inconsistent and incomplete. It is an interesting problem for further research to determine conditions under which such a description produces better results than a consistent and complete one.

Domain	Cpx/Rule	Attrs/Cpx	Values/Attr	Examples/Cpx
Lymphography	3	3.1	1.8	8
Breast cancer	20	3.9	1.7	5
Primary tumor	5.2	5.3	1.0	2.3

Table 3. Average complexity of AQ15's decision rules in the three medical domains, when no cover truncation mechanism was applied.

Domain	Tree pruning	Complexity Nodes	Leaves	Accuracy
Lymphography	no	38	22	76%
	yes	25	14	77%
Breast cancer	no	120	63	67%
	yes	16	9	72%
Primary tumor	no	188	90	41%
	yes	35	18	46%

Table 4. Average complexity and accuracy of decision trees built by ASSISTANT on 70% of examples, over 4 experiments. In all three domains the tree pruning mechanism reduced the complexity and increased the accuracy.

V CONCLUSION

A major contribution of the paper is to show that a relatively simple, attribute-based inductive learning method is able to produce decision rules of sufficiently high quality to be applicable to practical problems with noisy, overlapping and incompletely specified learning events. The AQ15 program has shown itself to be a powerful and versatile tool for experimenting with inductive knowledge acquisition in such problems. It produces decision rules which are easy to interpret and comprehend. The knowledge representation in the program is limited, however, to only attribute-based descriptions. For problems that require structural descriptions one may use a related program INDUCE2 [Hoff, Michalski & Stepp 83] or its incremental learning version INDUCE4 [Mehler, Bentrup & Riedsel 86]. A weakness of the experimental part of the paper is that the authors had no influence on the way the data were prepared for the experiments and the available data allowed us to test only a few of the features of AQ15.

Another major result is a demonstration that the knowledge reduction by truncating the covers may lead in some cases to a substantial reduction of the rule base without decreasing its performance accuracy. Further research will be required to find for any given domain a rule reduction criterion that leads to the best trade-off between accuracy and complexity of a rule base.

ACKNOWLEDGEMENTS

The authors thank Ivan Bratko and Igor Kononenko from the Faculty of Electrical Engineering at University of Ljubljana for collaboration and comments, and physicians Matjaz Zwitter and Milan Soklic from the Institute of Oncology at the University Medical Center in Ljubljana for providing medical data and helping to interpret them. We further acknowledge Gail Thornburg from the UI School of Library and Information Science and the AI Laboratory at the Dept. of Computer Science for her criticism and valuable suggestions.

REFERENCES

[1] Hong, J., Mozetic, I., Michalski, R.S. (1986). "AQ15: Incremental Learning of Attribute-Based Descriptions from Examples, the Method and User's Guide." Report ISG 86-5, UIUCDCS-F-86-949, Dept. of Computer Science, University of Illinois, Urbana.

[2] Hoff, W., Michalski, R.S., Stepp, R.E. (1983). "INDUCE.2: A Program for Learning Structural Descriptions from Examples." Report ISG 83-4, UIUCDCS-F-83-904, Dept. of Computer Science, University of Illinois, Urbana.

[3] Kononenko, I., Bratko, I., Roskar, E. (1986). "ASSISTANT: A System for Inductive Learning." *Informatica Journal*, Vol. 10, No. 1, (in Slovenian).

[4] Mehler, G., Bentrup, J., Riedesel J. (1986). "INDUCE.4: A Program for Incrementally Learning Structural Descriptions from Examples." Report in preparation, Dept. of Computer Science, University of Illinois, Urbana.

[5] Michalski, R.S. (1969). "On the Quasi-Minimal Solution of the General Covering Problem." *Proceedings of the V International Symposium on Information Processing (FCIP 69)*, Vol. A3 (Switching Circuits), Bled, Yugoslavia, pp. 125-128.

[6] Michalski, R.S. (1983). "Theory and Methodology of Machine Learning." In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), *Machine Learning - An Artificial Intelligence Approach*, Palo Alto: Tioga.

[7] Michalski, R.S. (1986a). "Concept Learning." To appear in *AI Encyclopedia*, John Wiley & Sons.

[8] Michalski, R.S. (1986b). "Two-tiered Concept Representation, Analogical Matching and Conceptual Cohesiveness." Invited paper for the *Workshop on Similarity and Analogy*, Allerton House, University of Illinois, June 12-14.

[9] Michalski, R.S., Chilausky, R.L. (1980). "Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis." *International Journal of Policy Analysis and Information Systems*, Vol. 4, No. 2, pp. 125-161.

[10] Michalski, R.S., Larson, J. (1975). "AQVAL/1 (AQ7) User's Guide and Program Description." Report No. 731, Dept. of Computer Science, University of Illinois, Urbana.

[11] Michalski, R.S., McCormick, B.H. (1971). "Interval Generalization of Switching Theory." Report No. 442, Dept. of Computer Science, University of Illinois, Urbana.

[12] Michalski, R.S., Mozetic, I., Hong, J. (1986). "The AQ15 Inductive Learning System: An Overview and Experiments." Report ISG 86-20, UIUCDCS-R-86-1260, Dept. of Computer Science, University of Illinois, Urbana.

[13] Michalski, R.S., Stepp, R.E. (1983). "Learning from Observations: Conceptual Clustering." In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), *Machine Learning - An Artificial Intelligence Approach*, Palo Alto: Tioga.

[14] Michalski, R.S., Winston, P.H. (1986). "Variable Precision Logic." AI memo No. 857, MIT, Cambridge. An extended version to appear in *AI Journal*.

[15] Mozetic, I. (1986). "Knowledge Extraction through Learning from Examples." In T.M. Mitchell, J.G. Carbonell, R.S. Michalski (Eds.), *Machine Learning: A Guide to Current Research*, Kluwer Academic Publishers.

[16] Quinlan, J.R. (1983). "Learning Efficient Classification Procedures and their Application to Chess End Games." In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), *Machine Learning - An Artificial Intelligence Approach*, Palo Alto: Tioga.

[17] Reinke, R.E., Michalski, R.S. (1986). "Incremental Learning of Decision Rules: A Method and Experimental Results." To appear in J.E. Hayes, D. Michie, J. Richards (Eds.), *Machine Intelligence 11*, Oxford University Press.