AN OUNCE OF KNOWLEDGE IS WORTH A TON OF DATA: Quantitative Studies of the Trade-Off between Expertise and Data based on Statistically Well-Founded Empirical Induction

Brian R Gaines
Knowledge Science Institute, University of Calgary
Calgary, Alberta, Canada T2N 1N4

ABSTRACT

There is currently a division in knowledge acquisition research and practice between techniques for the transfer of existing knowledge from human experts and those for the creation of new expertise through machine learning. This paper reports studies of the spectrum of trade-offs between these two extremes, measuring the amount of data required to attain knowledge through empirical induction given different forms and levels of expertise. This gives a principled economic evaluation of knowledge which can be used to guide knowledge acquisition theory and practice.

INTRODUCTION

There is currently a major paradigm split in knowledge acquisition research and practice between techniques for the transfer of existing knowledge from human experts and those for the creation of new expertise through machine learning (Gaines & Boose 1988, Boose & Gaines 1988). There is, however, a fundamental relation between the two paradigms in that existing expertise was at some time derived through empirical induction. There is also continuity between the two paradigms in that existing expertise may be partial, erroneous, and of various forms, such that it cannot completely replace empirical induction but may serve to to guide and expedite it (Gaines 1987).

For example human experts in a decision task might know:

1. **Minimal Rules**: a complete, minimal set of correct decision rules;
2. **Adequate Rules**: a set of decision rules that is complete in giving correct decisions but not minimal in containing redundant rules and references to irrelevant attributes;
3. **Critical Cases**: a critical set of cases described in terms of a minimal set of relevant attributes with correct decisions;
4. **Source of Cases**: a source of cases that contains such critical examples described in terms of a minimal set of relevant attributes with correct decisions;
5. **Irrelevant Attributes**: a source of cases as in 4 with correct decisions but described in terms of attributes among which are those relevant to the decision;
6. **Incorrect Decisions**: a source of cases as in 4 but with only a greater than chance probability of correct decisions;
7. **Irrelevant Attributes & Incorrect Decisions**: a source of cases as in 5 but with only a greater than chance probability of correct decisions.

This is a sequence of decreasing knowledge on the part of the human expert. It encompasses a range of situations met in practice, and it raises the question of how the amount of knowledge available from the expert affects the amount of data required for effective empirical induction. For case 1 no data is required for empirical induction since the correct answer is available. For case 7 the ‘expert’ has provided little except access to a source of data from which the correct answer might be derived. How much data is required for an optimal empirical induction procedure to derive 1 given 7, and how much less data is required for cases 2 through 6?

This paper reports some studies that give quantitative answers to these questions. The studies are empirical rather than analytic so that it is not apparent yet how the answers generalize to arbitrary situations. However, they establish some base-line data which is interesting in its own right, a possible guide to practitioners of knowledge acquisition, and a test case for potential analytic estimates of the ratios involved.

INDUCT: A STATISTICALLY WELL-FOUNDED EMPIRICAL INDUCTION PROCEDURE FOR DERIVING DECISION RULES FROM DATASETS

It is not the primary function of this paper to present yet another empirical induction algorithm. However, it is relevant to the assessment of the results to understand the method used to derive decision rules from data, and this section outlines the INDUCT algorithm, linking it to previous approaches, clarifying enhancements made and their roles in the studies, and comparing performance with the best known methods.

INDUCT is part of a knowledge acquisition tool KSS0 (Gaines 1987) which uses entity-attribute grid techniques (Shaw & Gaines 1983, Boose 1984) to elicit relevant attributes and critical entities from experts, and empirical induction based on these to build a knowledge base in terms of classes, objects, properties, values and methods (rules). The tool also accepts rules entered by experts and entity-attribute data from databases, so that its range of knowledge/data combinations encompasses all those listed in the previous section.

Cendrowska (1987) has shown that empirical induction through decision trees and direct conversion to rules, even with pruning, leads to rule sets that test the values of irrelevant attributes and are much larger than is necessary. Her PRISM algorithm goes from entity-attribute data direct to rules but does not address the problems of noisy data or missing values. Quinlan (1987) has developed extended pruning techniques in a way that are effective in coping with noisy data but still involve decision tree production with problems of irrelevant attributes and missing values.

INDUCT extends the PRISM algorithm to control direct rule generation through statistical tests that are effective in dealing with both noisy data and missing values. Figure 1 shows the basis for these statistical tests. Given a universe of entities, E, a target predicate, **Q**, and a set of possible test predicates of the form, **S**, on entities in E, use them to construct a set of rules from which the target predicate may be inferred given the values of the test predicates.

For the purposes of the statistical analysis the forms of **S** and **Q** do not matter. One may regard **S** as a *selector* choosing those e out of some subset of E for which to assert **Q**(e), and compare the selection process of the rule with that of random selection, asking "what is the probability that random selection of the same degree of generality would achieve the same accuracy or greater."

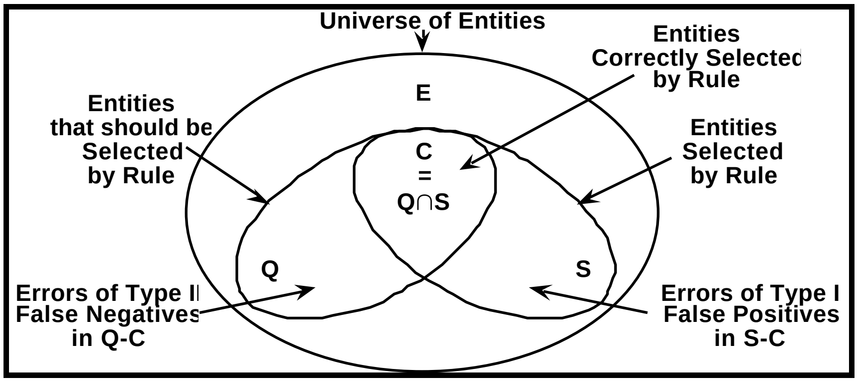
Universe of Entities
E
Entities Correctly Selectec by Rule
Entities that should be Selected by Rule
C = Q∩S
Entities Selected by Rule
Q
S
Errors of Type II False Negatives in Q-C
Errors of Type I False Positives in S-C

Fig. 1 Problem of empirical induction

This probability is easily calculated, let Q be the relevant entities in E for which **Q**(e) holds, S be the selected entities in E for which **S**(e) holds, C be the correct entities in E for which both **S**(e) and **Q**(e) hold:

$$Q \equiv \{e\!: e\quad E\quad \mathbf{Q}(e)\} \qquad S \equiv \{e\!: e\quad E\quad \mathbf{S}(e)\} \qquad C \equiv \{e\!: e\quad E\quad \mathbf{S}(e)\quad \mathbf{Q}(e)\} \tag{1}$$

Let the cardinalities of E, Q, S and C be e, q, s and c respectively. The probability of selecting an entity from E for which **Q** holds at random is p = q/e. The probability of selecting s and getting c or more correct at random is:

$$r = \sum_{i=c}^{s} {}^{s}C_{i}\, p^{i}\, (1-p)^{s-i} \tag{2}$$

The advantage of using r as a measure of the correctness of a rule is that is easily understood, as the probability that the rule could be this good at random, and that it involves no assumptions about the problem such as sampling distributions. Note that if c=s (all correct) then log(r) = s log(q/e) which seems to be the basis of 'information-theoretic' measures.

The probability that the performance of an isolated rule could be obtained by random selection is not adequate in itself to evaluate a rule that has itself been selected as best after a search through many possible rules. If n different rules have been searched then we may ask “what is the probability that the rule found by search achieved its performance by chance,” that is the probability that a rule will have been found that in itself has a probability of being selected at random of r. This probability is:

$$t = 1 - (1-r)^n \tag{3}$$

As one would expect this approximates to nr if this term is substantially less than 1.

To obtain a value for t one needs to estimate n, the number of rules in the search space, and this requires information about the forms of possible rules. If the rules are conjunctions of value tests of attributes as illustrated in (3) then n_1 can be estimated for rules with one clause, n_2 for rules with two or less clauses, and so on. For example if there are m attributes the i’th of which can be tested in m_i possible ways, then:

$$n_1 = \sum_{i=1}^{m} m_i \qquad\qquad n_2 = (n_1)^2 - \sum_{i=1}^{m} (m_i)^2 \tag{4}$$

Missing values are taken into account by assuming that they might have any value. When the selection of an entity is tested a missing value is assumed to have the required value for selection. In the statistics a selection based on missing values is allowed to contribute to false positives but not to correct positives. This has important consequence in knowledge acquisition since it allows the expert to enter conjunctive rules as if they were entities with missing values. INDUCT then generates the same or an equivalent smaller rule set. It is reasonable to test the consistency of an inductive procedure by requiring the rule-set produced by it to be ‘fixed-point’ if re-entered as data in this way.

INDUCT has been tested on a wide range of data sets in the literature together with many artificial data sets with known degrees of noise, missing values and irrelevant attributes, and found to perform consistently at least as well as the previously published best results. For example, with Quinlan’s (1987) very noisy data sets, Prob-Disj (disjunction of three terms in three attributes, one irrelevant attribute, 10% noise), and Digits (7-segment display, 10% noise in attributes, Bayes optimal solution 26% errors) the comparisons are:

Results on Prob-Disj Data (3-term disjunc

Data	Error Rate (Type I + Type II)	
	ID3 to Pruned Rules	INDUCT
Sample	4.2 rules	3 rules
Test1	10.0%	10.0%
Test2	10.0%	10.0%

Results on Digits Data

Data	Rules	Error Rate (Type I + Type II)	
		ID3 to Pruned Rules	INDUCT
Sample	10		31.05%
	16		26.8%
Test1	10		32.0%
	16	31.3%	29.0%
Test2	10		31.4%
	16	28.3%	27.8%

Fig. 2 Comparison of INDUCT and optimally-pruned ID3

THE TRADE-OF BETWEEN KNOWLEDGE AND DATA

Having established that INDUCT behaves well as an empirical induction algorithm, in particular, having high noise rejection, it is possible to use it to explore quantitatively the trade-offs between knowledge and data described in the Introduction. Cendrowska’s (1987) contact lens data used in her exposition of PRISM has been taken as a starting point since it is well-defined, previously analyzed and results in a range of rules of varying complexity, some of which are supported by a high proportion of the data set and others of which are supported by only single cases. The deterministic, complete data set involves 24 cases described in terms of 3 binary attributes, 1 ternary attribute and 1 ternary decision attribute. PRISM gives a solution based on 9 rules, but using default logic correct solutions with only 6 rules are available. On this data set INDUCT generates both the 9 rule and 6 rule solutions dependent on whether default logic is allowed (exception rules are given higher priority than partially-correct default rules).

This data set is used as a kernel from which to generate corrupted data with varying probabilities of errors and with varying numbers of irrelevant attributes. The generator selects an entity at random from the 24 cases, randomly changes the decision according to a prescribed probability, and adds a prescribed number of irrelevant binary attributes with random values. The corrupted data sets are run incrementally on INDUCT to determine the minimum amount of data necessary for a correct solution. This quantity is itself a random variable and 10 data sets generated with the same parameters are run to obtain more robust estimates of the data requirement.

Figure 3 shows the results obtained to date with a variation from 6 to over 1000 data items being needed to cover the spectrum of knowledge availability specified in cases 1 through 7 in the Introduction. It is interesting to note that high levels of noise can be tolerated—the expert by no means has to be 100% correct. Noisy irrelevant attributes cause similar effects to noise in the decision—this seems to validate claims that knowledge acquisition tools targeted on eliciting relevant attributes are in themselves worthwhile. There is strong interaction between errors and irrelevancy—much more data is required to eliminate both than either alone.

Knowledge	***Data Required***	
	Mean	**S.D.**
Exact rules	**6**	**0**
Critical cases	**18**	**0**
Correct cases	**90**	**43**
10% Errors	**123**	**49**
25% Errors	**326**	**159**
1 Irrelevant Att.	**160**	**77**
2 Irrelevant Atts.	**241**	**125**
5 Irrelevant Atts.	**641**	**352**
10% Err. + 1 Irr.Att	**1970**	**1046**

Fig. 3 Knowledge/data tradeoff in empirical induction

CONCLUSIONS

In theoretical terms, the approach taken in this paper seems to offer the possibility of developing a quantitative science of knowledge in terms of the amount of data reduction that knowledge buys us when carrying out empirical induction. This can be seen as a principled economic evaluation of the knowledge. The studies in this paper are empirical using a particular situation. It should be possible to obtain analytic results for more general cases.

In practical terms, the approach taken in this paper integrates into a coherent framework many of the different approaches being taken to knowledge acquisition, from those based on expert interviews to those based on empirical induction. The trade-off data is a guide to practitioners as to the appropriate approach and data requirements in their situations—it needs testing with other data sets but that is now a matter of sheer number crunching.

In design terms, the most important results obtained are that a single inductive algorithm is adequate to cover the complete spectrum of cases 1 through 7 in the Introduction, from rule simplification, to noise reduction and relevancy determination, and that it is possible to generate an integrated system that ranges across that spectrum, from the transfer of expertise from experts to the creation of equivalent expertise through empirical induction.

References

Boose, J.H. (1984). Personal construct theory and the transfer of human expertise. **Proceedings AAAI-84**, 27-33. California: American Association for Artificial Intelligence.

Boose, J.H. & Gaines, B.R., Eds. (1988). **Knowledge Acquisition Tools for Expert Systems**. London, Academic Press.

Cendrowska, J. (1987) An algorithm for inducing modular rules. **Int. J. Man-Machine Stud. 27** (4), 349-370.

Gaines, B.R. (1987). Rapid prototyping for expert systems. Oliff, M., Ed. **Proceedings of International Conference on Expert Systems and the Leading Edge in Production Planning and Control**. pp.213-241. University of South Carolina.

Gaines, B.R. & Boose, J.H., Eds. (1988). **Knowledge Acquisition for Knowledge-Based Systems**. London, Academic Press.

Quinlan, J.R. (1987) Simplifying decision trees. **Int. J. Man-Machine Studies 27** (3), 221-234 (September).

Shaw, M.L.G. & Gaines, B.R. (1983). A computer aid to knowledge engineering. **Proceedings of British Computer Society Conference on Expert Systems**, 263-271. Cambridge.