Prepublication draft of paper published in Machine Learning, 35, 5–21 (1998)

An Experimental Evaluation of Integrating Machine Learning with Knowledge Acquisition

GEOFFREY I. WEBB, JASON WELLS, AND ZIJIAN ZHENG {webb,wells,zijian}@deakin.edu.au
School of Computing and Mathematics, Deakin University, Geelong, Victoria 3217, Australia

Editor: Pat Langley

Abstract. Machine learning and knowledge acquisition from experts have distinct capabilities that appear to complement one another. We report a study that demonstrates the integration of these approaches can both improve the accuracy of the developed knowledge base and reduce development time. In addition, we found that users expected the expert systems created through the integrated approach to have higher accuracy than those created without machine learning and rated the integrated approach less difficult to use. They also provided favorable evaluations of both the specific integrated software, a system called *The Knowledge Factory*, and of the general value of machine learning for knowledge acquisition.

Keywords: Integrated learning and knowledge acquisition; classification learning; evaluation of knowledge acquisition techniques

1. Introduction

Machine learning[1] and knowledge acquisition from experts provide different and, on the face of it, complementary means of developing knowledge-based systems. The apparent manner in which the strengths of one match the weaknesses of the other has led to integration of the two approaches. This integration is expected to be synergistic in effect, the resulting combined approach being more effective than either of its components. However, although there have been case studies documenting successful applications of these integrated techniques (Buntine & Stirling, 1991; Morik, Wrobel, Kietz, & Emde, 1993; Nedellec, Correia, Ferreira, & Costa, 1994; Webb, 1996), no previous research has provided comparative evaluation of the relative merits of integrated approaches as opposed to either constituent approach on its own. In particular, it has not been demonstrated that integration of machine learning with knowledge acquisition can outperform with respect to any measure either of the original approaches alone.

Let us hasten to point out that we do not suggest that it is ever possible to perform machine learning in total isolation from a wider process of knowledge acquisition. To the contrary, machine learning cannot be performed without first selecting an appropriate class of model for the machine learning system to explore and specifying a suitable vocabulary and hence ontology for the domain. Indeed, it is credible that this is often the most significant part of the knowledge acquisition task. By the phrase "machine learning alone" we refer to the use of machine learning in a manner only loosely coupled with knowledge acquisition, in contrast to tightly integrated use of the technique.

In the research presented herein we sought to demonstrate that integrated use of machine learning can provide tangible benefits. In particular, we tested the hypotheses that the integration of machine learning with knowledge acquisition from experts can:

- produce more accurate expert systems than either constituent approach alone (the increased accuracy hypothesis);
- make knowledge acquisition less difficult than direct knowledge acquisition from experts without machine learning (the decreased difficulty hypothesis); and
- increase the developer's confidence in the accuracy of the resulting expert system (the increased confidence hypothesis).

We do not suggest that combining these techniques will always provide these benefits, but only that they will be apparent for appropriate knowledge acquisition tasks. Note that the decreased confidence hypothesis makes comparative predictions between knowledge acquisition from experts with and without machine learning. We do not predict that integrating machine learning with knowledge acquisition from experts will ever be less difficult than machine learning alone, as the former requires more input than the latter.

Ideally, evaluation of these hypotheses would examine true experts engaged in genuine knowledge acquisition tasks performed in real-world contexts. However, knowledge acquisition is a very expensive process. In most cases, the cost of applying multiple techniques to a single real-world knowledge acquisition problem would be prohibitive, let alone the cost of using each technique multiple times to enable statistical analysis of the observed differences. This may explain why previous evaluation has been restricted to case studies. Although such studies can demonstrate the capacity of a system to perform a specific task, they cannot provide comparative evaluation of alternative approaches. In view of these considerations, this research idealizes aspects of the knowledge acquisition process in order to perform controlled experimental comparisons of alternative techniques. In particular, we compare knowledge acquisition with and without integrated machine learning facilities.

2. Previous evaluation of integrated approaches

The literature contains a number of case studies demonstrating successful applications of techniques for integrating machine learning with knowledge acquisition from experts. Buntine and Stirling (1991) describe the development of an expert system for routing in the manufacture of coated steel products. While they do not use software that directly integrates machine learning with knowledge acquisition from experts, they describe a process that tightly couples the two. They emphasize the value of having an expert validate learned rules and place restrictions on the rules that are developed. They point to the importance of input from the expert in order that he accept the final expert system. While they state that 'a number of cases have been reported which demonstrate that the generic interactive induction

approach gives superior performance in real knowledge acquisition tasks to non-interactive induction and to knowledge acquisition by interview', the basis for this conclusion appears to be subjective judgment rather than experimental evaluation.

The knowledge acquisition tool MOBAL (Morik et al., 1993), and its forerunner BLIP (Morik, 1987), has been applied to a wide variety of knowledge acquisition tasks. These systems use first-order representations and they provide tools for specifying ontologies, learning rules from examples, revising rules using examples, and learning new predicates from examples. The user and learning systems collaborate to construct an expert system, with either party able to perform successive refinements to the evolving system.

Morik et al. (1993) present a case study in which MOBAL was used to acquire knowledge of basic German traffic laws. The resulting traffic-law expert system seeks to infer liability, likely fines, and whether a court appearance will ensue from specific traffic violation cases. The developers seeded the knowledge acquisition process by defining a set of 13 background rules, such as that parking is not permitted on a sidewalk, and then presented 12 example cases. A cycle of machine learning followed by refinement of the background knowledge ensued, until the developers were satisfied by the rules that the learning component inferred.

Another case study examined the use of MOBAL to create a prototype expert system for diagnosis of infantile jaundice (Morik et al., 1993). On the basis of experience applying the tools to develop this system, the authors conclude that the system's representation formalism is suitable for this domain; that the system's capabilities would usefully be extended to support revision of existing predicate definitions; and that the project demonstrated that the tools could be used as an expert system shell as well as for knowledge acquisition.

The Telecommunications Security Domain (Sommer, Morik, André, & Uszynski, 1994; Morik et al., 1993) is presented as a further study of knowledge acquisition using MOBAL. It covers the specification, validation and application of a security policy for communications network access control. The use of the tools led to the creation of a new concept, the *senior operator*, that had previously been missing from models of the domain. It also identified implicit policies that were pursued by network administrators and captured these policies in explicit rules.

Nedellec and Causse (1992) and Nedellec et al. (1994) provide brief descriptions of two case studies of the use of their tool APT. The application domains covered are design of loudspeakers and evaluation of commercial loan applications. In the first domain, few examples were available, so the developers used APT to assist the expert to generalize these examples and to 'elicit and correct' missing domain knowledge. Nedellec and Causse assert that 'the expert could easily evaluate the cases proposed by APT, and could understand when APT pointed to a deficiency in the domain theory' and that 'the knowledge acquisition principles of APT help the expert to identify missing or incorrect knowledge and to integrate the modifications in an efficient way'. For the commercial loans domain, Nedellec and Causse used APT to refine an existing expert system, producing a system that was one fifth the size of the original and consequently more efficient. It also provided structure to the domain knowledge that was missing from the original. This case study

provides a comparison between knowledge acquisition with and without integrated use of machine learning. However, the integrated approach had a considerable advantage as the project took a system developed by knowledge acquisition alone and examined the effects of using machine learning for its further refinement. It remains unclear how refining the initial expert system using conventional knowledge acquisition without machine learning would have fared had equivalent effort been expended.

Webb (1996) describes a case study in which undergraduate computer science students used The Knowledge Factory to produce expert systems for an artificial medical domain. Evaluation of questionnaires administered on completion of the project revealed that the students found the system easy to use, regarded the system to be a valuable tool, and believed that machine learning was useful for knowledge acquisition. Webb concluded that 'these results can be considered as a proof-of-concept for the proposition that non-knowledge-engineers can readily collaborate with a machine learning system to develop expert systems'. Again, however, the study provided no comparative evaluation of the integrated approach against any alternative. Techniques that integrate machine learning with knowledge acquisition from experts have been demonstrated to work, but do they work better than, or even as well as, the alternatives?

3. Experimental design

We wished to explore this question by comparing the integrated use of machine learning during knowledge acquisition with each constituent approach in isolation. Only if the integrated approach demonstrably outperforms each of its constituent approaches can it credibly be claimed that the integration provides benefit per se.

Matched-pairs experimental designs allow more powerful comparison of treatments than independent sample comparisons. In consequence, they should be used where possible. We did not believe we could adequately match experts, however, and hence used a less powerful independent samples design to compare knowledge acquisition by experts with and without access to machine learning. In past experiments we have sought to overcome this problem by using the same expert in both treatments, allowing us to match each expert with himself, but it is also necessary to match the experimental units with respect to knowledge acquisition task. A single expert performing two different tasks would not adequately control all factors other than the experimental manipulation. We previously attempted this by disguising a single knowledge acquisition task as two different tasks (Webb & Wells, 1996). However, such an attempt can always be criticised on the grounds that the expert's view of the task is an important aspect of a knowledge acquisition scenario, and even apparently superficial differences in the expert's perceptions may affect their behavior.

To maximize the statistical power of the analyses, given the small numbers of subjects, we tested each subject twice, once for each treatment excluding learning alone, for which subjects were not required. To minimize the introduction of experimental confounds through order and practice effects, and due to differences

between the two tasks, we defined two knowledge acquisition tasks and evenly divided subjects between each treatment for each task. Each subject received one treatment for one task in the first session and then the other treatment and task in the second session. To minimize order effects and confounds introduced by task differences, we assigned the integrated version as the first task for half of the subjects while we assigned it as the second task for the other half. We randomized this assignment using a random number generator. By randomly assigning subjects to treatments, all four (2×2) combinations of treatments and tasks were covered.

In contrast to the need for unmatched comparisons between the integrated and knowledge acquisition alone treatments, matched-pair comparisons were possible between the integrated and learning alone treatments. For the latter treatment, we used the machine learning system to induce rules from the same training data we provided to a subject when they were using the integrated software. We then evaluated these rules against the matching test data.

Creating these three treatments made it possible to test the increased accuracy hypothesis by examining the performance of expert systems created under each condition. We randomly selected training and test data for each subject for each task. The training data were available during knowledge acquisition but the test data were withheld. We then evaluated the expert system by applying it to the test data. Different random selections of training and test cases were performed for each subject, in order to minimize the possibility that atypical training/test splits would unduly affect the experimental results. As noted above, the training and test data used in the integrated treatment were also used in the learning alone treatment, permitting matched-pairs analysis.

The remaining two hypotheses relate to less tangible aspects of performance and required more indirect evaluation. One indirect effect of reducing knowledge acquisition difficulty might be to reduce knowledge acquisition time. A less difficult task will usually be completed more swiftly. Hence, the relative times taken to complete a task under each treatment can be taken as an indirect source of evidence with respect to the decreased difficulty hypothesis. Significant reductions would seem to support the hypothesis, although they would not be conclusive. A second indicator of task difficulty is the expert's subjective judgment. A questionnaire was designed to elicit such judgments from the subjects. For ease of presentation, the questionnaire design and results are discussed together, below.

The increased confidence hypothesis related directly to subjective judgment, so we extended the questionnaire to canvas the subject's confidence in the accuracy of the expert systems created under each treatment. As straight machine learning does not directly involve the subject in the development of the expert system, it appeared appropriate only to make this comparison between the treatments in which the subjects would be active participants, the integrated and knowledge acquisition alone treatments.

While not pertaining to our hypotheses, we also measured the relative complexity of the knowledge bases developed in the belief that this may interest some researchers and practitioners.

We needed to create two knowledge acquisition tasks to which the three treatments could be applied. To facilitate fair comparison between treatments and to keep the complexity of the experimental task within manageable bounds, we limited the tasks to those aspects of knowledge acquisition to which machine learning can be directly applied without the need for further involvement from an expert. To this end, we selected data sets from the UCI repository of machine learning data sets (Merz & Murphy, 1997) to form the core of the knowledge acquisition tasks. This meant that the tasks were constrained as the selection of model types, vocabularies, and ontologies were already completed.

We needed, however, to employ experts for the integrated and knowledge acquisition alone treatments. To ensure that levels of expertise were controlled, we "created" these experts by providing the subjects with expertise in the domains.

We acknowledge that this experimental design results in simple artificial knowledge acquisition tasks that have abstracted out important elements of full-scale real-world knowledge acquisition projects. Such simplification is an inevitable price of controlled experimentation and it is standard practice in the social sciences. Factoring out possible confounds is important if we are to systematically evaluate hypotheses. We believe that we have managed to retain non-trivial aspects of knowledge acquisition in our design. Subjects are provided with expertise in a domain and tools with which to express, explore, and evaluate models that capture that expertise. We have carefully controlled the scenario so that the only difference between those tools is that one includes tightly integrated machine learning facilities and the other does not.

Section 5 provides the details of how we implemented this design. Before presenting these details, in the next section we discuss the knowledge acquisition software used in the experiments.

4. The Knowledge Factory

This research was part of a project that developed The Knowledge Factory system. It was therefore natural that we should use this system for the study. We present here a brief description of this system and relate it to other software for integrating machine learning with knowledge acquisition from experts. More details about the system are available elsewhere (Webb, 1992, 1996; Webb & Wells, 1995).

The Knowledge Factory is an interactive environment that was developed with the intention of enabling a domain expert to collaborate with a machine learning system throughout the knowledge acquisition and maintenance process. Most approaches to integrating machine learning and knowledge acquisition from experts require the involvement of a trained knowledge engineer (Attar Software, 1989; Buntine & Stirling, 1991; Davis & Lenat, 1982; De Raedt, 1992; Gams, Drobnič, & Karba, 1996; Morik et al., 1993; Nedellec & Causse, 1992; O'Neil & Pearson, 1987; Schmalhofer & Tschaitschian, 1995; Smith, Winston, Mitchell, & Buchanan, 1985; Tecuci & Kodratoff, 1990; Wilkins, 1988). Like the approach of Tecuci (1995), The Knowledge Factory is distinguished by being designed for direct use by experts with minimal training or experience in knowledge engineering. The system also differs

from a number of knowledge elicitation systems designed for direct use by experts (Boose, 1986; Compton, Edwards, Srinivasan, Malor, Preston, Kang, & Lazarus, 1992), not only by the provision of machine learning facilities, but also by not relying upon the expert to give suitable solutions for all cases that are encountered.

The Knowledge Factory employs simple knowledge representation schemes in order to accommodate the target user group, as many users have great difficulty using the first-order representations commonly used by knowledge acquisition environments (e.g., Kodratoff & Vrain, 1993). In consequence, we restricted the knowledge representation scheme to flat attribute-value classification rules and the knowledge base to a set of production rules. Moreover, the antecedent of a rule consists of tests on attribute values and the consequent is a simple classification statement. All rules directly relate input attributes to an output class. This simple attribute-value representation contrasts with the first-order representations used by most recent integrated systems (De Raedt, 1992; Morik et al., 1993; Nedellec & Causse, 1992; Schmalhofer & Tschaitschian, 1995; Tecuci & Kodratoff, 1990; Wilkins, 1988).

Due to the needs of the target user group, we have also kept the interface simple. An approach called *case-based communication* motivates the primary mechanisms for interaction between the expert and the machine learning system. The system communicates the support for the rules that it develops by displaying the example cases that a rule covers correctly, covers incorrectly, or fails to cover, as well as the cases that the rule set as a whole classifies correctly, classifies incorrectly, or leaves unclassified. The expert can critique rules by providing complete or partial counterexamples, and can explore alternatives to current rules by specifying cases that a rule should or should not cover. In addition, the user has facilities for directly editing rules and example cases and for modifying the set of attributes with which example cases are specified.

The machine learning facilities support induction of new rules and inductive refinement of existing rules. Machine learning can be applied at any time to create or revise a complete rule set or the subset of rules that relate to a single class only. During inductive refinement, the user can restrict the types of modification possible for each existing rule. The machine learning algorithm, DLGref2 (Webb, 1993), adds rule refinement capabilities to DLG (Webb & Agar, 1992), which is in turn a variant of Michalski's (1983) AQ. DLGref2 seeks to modify an existing set of rules the least amount necessary to optimize a user specified preference criterion. When there are no pre-existing rules, DLGref2 is equivalent to DLG.

5. Experimental method

We developed two versions of The Knowledge Factory, one containing the machine learning facilities and one with these facilities excised. The former was used for the integrated treatment and the latter was used for the knowledge acquisition alone treatment. For the learning alone treatment, the experimenters applied the machine learning facilities of The Knowledge Factory to the learning tasks.

Note that the version of the software without machine learning capabilities is not a straw man. Even with the learning facilities removed, The Knowledge Factory is still a fully functional knowledge acquisition environment. The system contains both extensive facilities for specifying and editing rules and for evaluating the performance of those rules on example data.

The experiments were conducted as part of an assignment for a third-year undergraduate university course in computing. The use of undergraduate computing students with minimum knowledge acquisition training and no knowledge acquisition experience seemed appropriate, as the tool is intended for users with little training in knowledge engineering.

5.1. The knowledge acquisition tasks

All 18 students in the third year unit *Artificial Intelligence and Expert Systems* at Deakin University were given an assignment that involved two knowledge acquisition tasks. The student body comprised both Information Systems and Software Development students. All students involved were asked whether they would consent to take part in a research study and were told that they could withdraw their consent at any stage during the experiment. One student exercised the option of not participating, leaving 17 subjects.

The study commenced in the sixth week of the unit. Up to that point, the students had been exposed to overviews of knowledge acquisition principles and techniques, and they had been taught programming in the CLIPS expert system language. During the study, the students received further lectures and laboratory sessions on CLIPS programming and two discursive lectures on knowledge acquisition principles and techniques. Thus, while having good computer skills, the subjects were, at best, novice knowledge engineers.

The Glass and Soybean Large data sets from the UCI repository were selected to form the tasks. We chose these data sets to provide a mix of different characteristics. Of the two possible class variables for the Glass data set, we used the one that defines three classes: Float, Not_float, and Other. The Soybean Large data set defines a task with 19 classes (different diseases). The Glass data set has nine real valued attributes while the Soybean Large data set has 35 categorical attributes. The Glass data set has 214 cases while Soybean Large has 683 instances.

To provide subjects with expertise in the domains, we ran C4.5rules (Quinlan, 1993) to create sets of rules and then presented these rules to the subjects, providing them with background knowledge about the domain. C4.5rules was used in order to provide different insights and analysis from those provided by The Knowledge Factory. To simulate different types of expert knowledge, we used different procedures to generate these rules for each data set.

For the Glass domain, background knowledge was generated by running C4.5rules on the entire data set and then extracting all the rules for the class Other. Only the rules for this single class were presented to the subjects. This simulates a situation where the expert has an accurate and well-defined procedure for identifying cases of one class but not the others.

For Soybean Large, one half of the data (342 cases) was randomly selected for background knowledge generation. As this selection process was independent of that used to select training and test cases during testing, described below, some of the selected cases belonged to a subject's training set and others to the test set. C4.5rules was applied to this random selection of cases to develop a set of rules. This simulates a situation where the expert has considerable insight into the classification task, but does not have the ability to perfectly classify all possible cases. Whereas the background rules for Glass enabled correct classification of both training and test cases for one class only, the background rules for Soybean Large enabled reasonably accurate, but not perfect, classification for all classes. The accuracy of the background rules when applied to the entire Soybean Large data set was 86.1%.

The knowledge acquisition tasks were defined by random selection of training and test cases. For Glass, each subject was allocated 172 training and 42 test cases. For Soybean Large, each subject was allocated 342 training and 341 test cases.

All subjects participated in three laboratory sessions of three hours each at one week intervals. Table 1 presents the instructions given to each subject. The first session provided training in use of the software and an introduction to the type of task that they were to perform. The subjects were provided with both Glass and Soybean Large training data, on which the software was demonstrated and on which they could practice. In particular, they were shown how the rules developed using training data could be evaluated with respect to test data. They were informed that such evaluation would be used to grade the expert systems that they developed in the subsequent two sessions. Subjects received different sets of training data in the training and experimental sessions and access to the data provided in the training session was restricted to the duration of that session.

It is possible and desirable that the subjects gained some expertise in glass fragment analysis and soybean disease diagnosis during the training session, but it is probable that the subjects remembered little domain-specific knowledge over the one week interval between sessions, other than the background rules with which they were provided. Subjects were not provided with access to their test data. Evaluation was performed by the experimenters after the laboratory session.

The software, manuals, and data were given to the subjects on a computer disk. The task was performed in a supervised computing laboratory environment. Subjects were able to ask questions of the experimenters at any stage during the experiment, but responses were restricted to details directly relating to how to operate the software. Other than this, the only assistance that the subjects obtained was in the form of access to the system's help facilities and the user manual.

5.2. Software employed

The Knowledge Factory operates within the Macintosh software environment. Previous experience with student use of The Knowledge Factory software had shown that there was a tendency for students to explore the full range of features that it provided, including multiple modes of machine learning and multiple modes of

Table 1. Instructions given to subjects in the study.

SCC376 Assignment 1 1997
To be completed in laboratory sessions in the weeks starting April 21 and April 28.
20 marks (10 marks for each part)

This assignment places you in the role of an expert system developer. You will have two knowledge acquisition tasks. Each is to be performed with a different version of The Knowledge Factory knowledge acquisition software. One version includes machine learning facilities and the other does not. Your assignment disks each provide

- a copy of The Knowledge Factory (TKF.Induction_On or TKF.Induction_Off);
- a The Knowledge Factory project file containing 200 example cases (<id>.prj)
- a text file detailing your knowledge of subject matter (<id>.briefing)
- a manual in Microsoft Word format (Manual.Induction_On or Manual.Induction_Off).

Your task is to use that copy of The Knowledge Factory to create rules for that domain.
The two tasks are soybean disease diagnosis (Part A) and glass fragment analysis (Part B). You will be provided with a summary of knowledge about these two tasks along with sets of case histories of past cases.
Assignments will be marked on their predictive accuracy in classifying previously unseen cases. The 'correct' rules for each task will vary from student to student, so there is no benefit in sharing data or rules with other students.
Different students will be randomly assigned different versions of The Knowledge Factory for each task to allow us to study the relative power of different aspects of the system. Marks will be standardized for all students with each version of the system for each task so that no student is disadvantaged. The results of this comparison will be discussed in class.
We wish to use outcomes of this assignment for research purposes to evaluate the strengths and weaknesses of the different versions of The Knowledge Factory that are used. To this end you will be given a consent form. If you do not wish to have your project included in this research, tick the box labeled "do not consent". Otherwise tick the box labeled "hereby consent". Those that do not participate in the research project will be in no way penalized. Those that do participate will not be personally identified for research purposes and will contribute to research on knowledge acquisition that will be of benefit to the class and to the expert systems development community.
The materials for these tasks will be distributed in the laboratory sessions. You will be marked on the expert systems on disk when the session is completed. A questionnaire will be distributed at the end of the last laboratory session, which participants will be asked to complete.
If you have any questions please contact Zijian Zheng (room SD108, telephone 5227 1325).
Thank you in anticipation for your cooperation.

rule interpretation (Webb, 1996). As these capabilities did not bear directly upon the issues to be explored by this study, they were disabled. The default machine learning and rule interpretation settings were employed, with one exception.

By default, The Knowledge Factory applies rules in a mode that lets the system withhold decisions. This outcome occurs when no rule covers a case or when multiple rules for different classes cover a case. Such results make it extremely difficult to compare the performance of alternative expert systems, as there is no definitive manner in which to compare a system that achieves an accuracy of $x_1\%$ on $y_1\%$ of cases for which it reaches a conclusion with a system that achieves $x_2\%$ accuracy on $y_2\%$ of cases.

To obviate this problem, The Knowledge Factory was set to a mode in which, when no rule applied to a case, it assigned the most common class from the training set,

and when multiple rules covered a case, it assigned the class predicted by the highest quality rule (in terms of performance on the training set). For this experiment, the quality of a rule was judged by the function

$$quality = \begin{cases} -1 & \text{if } n > 0 \\ p & \text{otherwise} \end{cases} \quad (1)$$

where p is the number of cases correctly classified by the rule and n is the number of cases incorrectly classified. With this evaluation function, the specific-to-general search used in this learning algorithm avoids rules that cover any negative cases. As a consequence, there is no need to distinguish between the quality of alternative rules that cover negative cases.

Further features of the system that did not directly bear upon the experimental question but that had potential to seriously degrade performance if misused were also disabled. These were:

- all facilities for adding, deleting, or otherwise transforming attributes, as subjects had access to no source of knowledge that could warrant such actions.

- the ability to generate new cases, as subjects had no knowledge by which to generate new reliable example cases.

- the ability to load from external files either additional example cases or sets of rules, as these could not be used in a sensible manner within the scope of the experiment.

- the facility for deleting example cases, as subjects were informed that all example cases were accurate and hence had no basis for sensible deletion.

- mechanisms for dividing the example cases into training and test sets, as the number of example cases made available to the students was too small for this facility to be useful.

In addition, to prevent subjects from exchanging data between versions of the system or using other data analysis tools, the students were prevented from outputting the data in any form other than as a project file, the system's internal format for data representation.

To simplify the task of tracking progress, subjects were presented with a computer disk containing the appropriate version of the system along with a project file preloaded with the training data. The software was modified to require the system to be run from that disk and only on the original project file (although that file could be updated by The Knowledge Factory under the user's direction).

The software was also modified to ensure that projects saved by one version of the system could not be input into another. This prevented subjects in one condition from obtaining and using a copy of the software for the other condition.

5.3. Experimental manipulation

Two versions of the software were created. The version enabled with machine learning had the full functionality of The Knowledge Factory software other than the disabled features noted above. The knowledge acquisition alone version was identical to the integrated version, except that four commands were disabled:

- Develop New Rules deletes any existing rules and then applies the DLG machine learning algorithm (Webb & Agar, 1992) to the training examples to form a new set of rules.

- Revise Current Rule Set applies DLGref2 (Webb, 1993) to refine the current set of rules. This inductive refinement algorithm seeks to modify each of the existing rules the least amount necessary in order to optimize the preference criterion, as defined by equation 1. The user can specify that selected rules not be modified in this process. After all existing rules have been processed, new rules are added to the rule set to cover any example cases not covered by the modified rule set.

- Revise Rules For Current Decision is identical to Revise Current Rule Set, except that only rules for the class of the currently selected rule are modified or added to the rule set.

- Form Alternative Rules takes an existing rule and presents a set of alternative rules that correctly classify all cases correctly classified by the original rule and incorrectly classify no example case that was not incorrectly classified by the original rule.

We should emphasize that, although the knowledge acquisition alone version of the software did not contain the machine learning capabilities described above, it still retained a comprehensive set of facilities for rule specification, editing, and evaluation.

5.4. Likely knowledge acquisition processes

As experienced users of the software, we would have approached the subjects' tasks in the following manner. With the integrated system, we would:

1. specify rules for existing domain knowledge;

2. apply the machine learning facilities to refine the current rules;

3. evaluate the resulting rule set in the light of both the original domain knowledge and case-based evaluation of rule performance;

4. if deficiencies are detected, modify the rule set by direct editing or using the example-based editing facilities, then return to step 2.

Table 2. Mean and standard deviation for predictive accuracy.

Data set	Integrated System	KA Alone	Learning Alone
Soybean Large	88.5 ± 4.4	84.4 ± 3.6	87.8 ± 1.6
Glass	81.3 ± 7.7	59.0 ± 17.3	79.2 ± 8.0

We would have used the same procedure with the knowledge acquisition alone system but for step 2. During the training session we used the full version of this process with the integrated system to demonstrate the software. However, while it was used in the training session, it was not explicitly described to the subjects. Subjects were free to employ the software as they saw fit.

6. Evaluation of expert system quality and acquisition difficulty

Seventeen students consented to participate at the commencement of the study and none withdrew thereafter. The expert systems that these subjects developed were compared on predictive accuracy and number of rules. Knowledge acquisition time was also compared.

6.1. Predictive accuracy

Table 2 presents the mean predictive accuracy obtained for each treatment. The mean accuracy for the integrated treatment is significantly higher than that for knowledge acquisition alone in each each domain (one-tailed two-sample t tests; Soybean Large: $t = 2.1$, $p = 0.026$; Glass: $t = 3.35$, $p = 0.001$).

When compared with the predictive accuracy obtained by the subjects in the integrated condition, both mean accuracy results for learning alone are lower than the corresponding means obtained through the use of machine learning with knowledge acquisition from experts. One-tailed matched-pairs t tests reveal that one of these differences is significant at the 0.05 level (Glass: $t = 2.5$, $p = 0.021$) but the other is not (Soybean Large: $t = 0.9$, $p = 0.279$). It should be noted, however, that the power of these comparisons is low (only eight and nine pairs being involved, respectively) and hence that the failure to obtain a significant difference for the Soybean Large data provides only weak evidence that there was no advantage for the integration of machine learning with knowledge acquisition for this domain.

To assess the quality of the background knowledge with which subjects were provided, it is valuable to consider the accuracy of the rules supplied for this purpose when applied to the subjects' test sets. On the training and test sets in the integrated condition, for Soybean Large the mean accuracy was 86.8 ± 1.4 and for Glass it was 64.3 ± 6.1. For the knowledge acquisition alone training and test sets, the mean accuracies were 85.8 ± 1.8 for Soybean Large and 62.4 ± 5.4 for Glass. These variations between treatments are to be expected for such small sample sizes and do not represent a statistically significant difference (two-tailed two-sample t tests; Soybean Large: $t = 1.3, p = 0.211$; Glass: $t = 0.7, p = 0.518$). Comparing

Table 3. Mean and standard deviation for time in minutes.

Data set	Integrated System	KA Alone
Soybean Large	73 ± 45	131 ± 19
Glass	16 ± 19	115 ± 38

the performance of the background knowledge rules with the accuracy of the rules produced in the corresponding integrated and knowledge acquisition alone treatments, the only significant difference is an increase in accuracy for the integrated treatment on the Glass data (one-tailed matched-pairs t tests; integrated Soybean Large $t = 1.1, p = 0.153$; integrated Glass: $t = 9.6, p < 0.001$; knowledge acquisition alone Soybean Large: $t = 1.0, p = 0.186$; knowledge acquisition alone Glass: $t = 1.2, p = 0.134$). While the decreases in accuracy from the background knowledge rules to the rules formulated by the subjects in the knowledge acquisition alone treatment are not significant, the power of the tests is low given the small sample sizes. It seems credible that subjects have in general failed to capture all information available to them in the models they created.

In summary, the integrated treatment has created expert systems for each domain that have higher mean accuracy than those created by the knowledge acquisition alone treatment, learning alone, or the background rules with which subjects were provided. While these differences were not significant for the Soybean Large domain, they were for the Glass domain, providing support for our increased accuracy hypothesis.

6.2. Knowledge acquisition time

The second major variable analyzed was knowledge acquisition time. Recall that we used this factor as an indirect measure of knowledge acquisition difficulty. We predicted that integrated knowledge acquisition would be less difficult than knowledge acquisition without machine learning and hence that the subjects in the integrated condition would take less time to complete their projects than those in the knowledge acquisition alone condition. The mean knowledge-acquisition times in minutes are presented in Table 3. One-tailed two-sample t tests revealed significant differences between integrated and knowledge acquisition alone treatments for each domain (Soybean Large: $t = 3.3$, $p = 0.002$; Glass: $t = 1.9$, $p = 0.037$). The incorporation of machine learning significantly reduced acquisition time. Indeed, for the second task, when the subjects were more experienced in the use of the software and the performance of these types of task, the average knowledge-acquisition time with the use of machine learning was less than one tenth of the time without.

Times were not recorded for the learning alone treatment, as this did not relate to the experimental hypotheses. However, as the machine learning facilities of The Knowledge Factory take less than a minute to complete an expert system for each full data set from which the training sets were drawn, it is safe to conclude that learning alone is substantially faster than either of the other approaches.

Table 4. Mean and standard deviation for number of rules.

Data set	Integrated System	KA Alone	Learning Alone
Soybean Large	46.9 ± 11.9	44.8 ± 25.6	35.3 ± 2.3
Glass	19.0 ± 3.0	13.7 ± 8.4	17.1 ± 1.6

6.3. Complexity

Although no predictions were made about the complexity of the rule sets formed by each treatment, results are presented here for the interested. We recorded the number of rules for each expert system developed and present the mean and standard deviation for each treatment in Table 4. Two-tailed two-sample t tests revealed no significant differences between integrated and knowledge acquisition alone treatments for either domain (Soybean Large: $t = 0.225$, $p = 0.832$; Glass: $t = 1.7$, $p = 0.222$). One-tailed matched-pairs t tests showed that the increase in complexity from learning alone to integrated was significant for Soybean large but not Glass (Soybean Large: $t = -3.0$, $p = 0.016$; Glass: $t = -2.0$, $p = 0.089$). Although most of these results are not significant, little comfort can be drawn therefrom, as the small number of subjects meant the analyses had low statistical power. It remains credible that the increase in accuracy resulting from combination of machine learning with knowledge acquisition from experts is matched by correspondingly more complex expert systems. However, this increased complexity may be needed to obtain the improved accuracy, rather than being a direct product of the integration process.

7. Questionnaire

The questionnaire was completed by the subjects at the end of the last laboratory session. Table 5 presents the questions asked and mean ratings provided by subjects in response. All ratings used a scale of 1 to 5, with 1 denoting *not at all* and 5 denoting *very*. Questions 1a to 4a did not appear on the questionnaire, but rather were derived from responses to questions 3 to 6 as described below.

7.1. Questionnaire design

We designed this questionnaire to examine a number of issues. The first four questions relate to the hypothesis that the integrated approach would be less difficult than knowledge acquisition alone. To minimize confounds from expectancy effects, whereby subjects' responses are influenced by their beliefs about the experimenters' expectations, questions comparing the integrated and knowledge acquisition alone treatments were either measured indirectly or cross-validated via indirect assessment. Questions 1 and 2 were cross-validated by the indirect questions 1a and 2a which were derived by reinterpretation of subjects' answers to questions 3 and 4. For subjects given the integrated software for the soybean disease diagnosis task,

Table 5. Questionnaire given to subjects after the study.

No.	Question	Rating
1	How difficult was it to use "TKF.Induction_On"?	1.8
2	How difficult was it to use "TKF.Induction_Off"?	3.4
3	How difficult was it to create an expert system for soybean disease diagnosis?	3.2
4	How difficult was it to create an expert system for glass fragment analysis?	2.7
5	How accurate do you think the expert system you created for soybean disease diagnosis will be when applied to the additional unseen evaluation cases?	3.3
6	How accurate do you think the expert system you created for glass fragment analysis will be when applied to the additional unseen evaluation cases?	3.4
7	Do you think that "TKF.Induction_On" is a useful tool for building expert systems?	4.5
8	Do you think that "TKF.Induction_Off" is a useful tool for building expert systems?	3.3
9	How valuable do you think machine learning is for knowledge acquisition?	4.1
1a	How difficult was it to create an expert system for the task for which you used TKF_induction_on?	2.1
2a	How difficult was it to create an expert system for the task for which you used TKF_induction_off?	3.8
3a	How accurate do you think the expert system you created using TKF_induction_on will be when applied to the additional unseen evaluation cases?	3.8
4a	How accurate do you think the expert system you created using TKF_induction_off will be when applied to the additional unseen evaluation cases?	2.9

Note. The names "TKF.Induction_On" and "TKF.Induction_Off" referred to the integrated and knowledge acquisition alone versions of The Knowledge Factory software, respectively.

the results for 1a and 2a are the responses for questions 3 and 4, respectively. For the remaining subjects, the results for 1a and 2a are the responses for 4 and 3, respectively. For example, a subject with learning enabled software for the soybean domain was given the same rating for the implicit question 1a, *How difficult was it to create an expert system for the task for which you used TKF_induction_on?*, as they provided for question 3, *How difficult was it to create an expert system for soybean disease diagnosis?*

To evaluate the increased confidence hypothesis, questions 5 and 6 were designed to examine the effect of the system employed on the subject's perception of the quality of the knowledge base developed. To reduce expectancy effects, the systems were referred to by task rather than by system. However, the subjects' answers to these questions were reinterpreted as 3a and 4a. For a subject given integrated software for the soybean disease diagnosis task, the result for 3a was the response to question 5 and the result for 4a was the response to question 6. For the remaining subjects these pairings were reversed.

While not directly pertaining to the experimental hypotheses, questions were added seeking to explore subject's judgments about the value of the integrated approach. Questions 7 and 8 were designed to evaluate the subject's perception of the relative usefulness of the two versions of the system. Question 9 was designed to elicit the subject's perception, after using each version of the software, of the value

of the main distinguishing feature between the two versions. It was not apparent how expectancy effects might be minimized for these questions.

7.2. Questionnaire results

Because we hypothesized that subjects would find it easier to develop an expert system with the aid of machine learning, we predicted that the responses would be lower for question 1 than 2. One-tailed matched-pairs t tests supported this prediction ($t = -5.5$, $p < 0.001$). Questions 1a and 2a were designed to cross-validate the results for questions 1 and 2. A one-tailed matched-pairs t test confirmed the analogous prediction that the mean answer for 1a should be lower than that for 2a ($t = -4.9$, $p < 0.001$).

No predictions were made with respect to differences between questions 3 and 4 because the relative difficulty of the two tasks was unknown. Two-tailed matched-pairs t tests showed no significant differences ($t = 1.0$, $p = 0.355$), so we cannot conclude that subjects felt either problem to be more difficult.

Subjects were expected to anticipate higher predictive accuracy when using machine learning (3a) than when not (4a) and this prediction was confirmed by a one-tailed matched-pairs t test ($t = 2.7$, $p = 0.009$)[2]. No prediction was made with respect to whether the subjects would expect a difference in predictive accuracy between domains (questions 5 and 6). A two-tailed matched-pairs t test of the sixteen pairs of answers to these questions showed no significant differences ($t = 0.2$, $p = 0.887$), providing no evidence that subjects expected higher predictive accuracy for either domain.

However, we did expect subjects to regard the version of the software that provided machine learning facilities as more useful than the version that did not (questions 7 and 8). This prediction was confirmed by one-tailed matched-pairs t tests ($t = 4.6$, $p < 0.001$). Similarly, subjects were expected to provide high ratings for the value of machine learning for knowledge acquisition (question 9), and a one-tailed t test shows that the mean responses were significantly higher than 3, the middle value ($t = 7.7$, $p = 0.005$).

These results show that the subjects believed the machine learning facilities to be useful, found the knowledge acquisition process easier when the machine learning facilities were available, and had greater confidence in the expert systems developed with the aid of machine learning.

8. Discussion

The subjects in this experiment had minimal expertise in knowledge engineering and were given limited training in the use of the software. This was intended both to prevent the experimenters from unduly guiding the subjects and hence confounding the results and to simulate the use of the software by domain experts rather than knowledge engineers. We acknowledge, however, that the third year computing students employed in this study have more sophisticated computer awareness than the typical domain expert and that this may have influenced results. The limits

of the subjects' domain expertise should also be noted. Providing subjects with a set of classification rules can be expected to produce only a low fidelity simulation of real-world expertise. Domain experts utilize many types of knowledge and it is debatable whether any resemble simple classification rules. The motivations of real experts also differ from those of our subjects, for whom the primary motivation was to maximize student grades. Further, key steps of the knowledge acquisition task were done for the subjects, namely, the selection of the type of model to form and the definition of a vocabulary (here the attributes and their values) and hence ontology.

Notwithstanding these caveats, the subjects were provided with some form of expertise in each domain, they were motivated to maximize the predictive accuracy of the expert systems that they created, and the experiment systematically evaluated the relative efficacy of knowledge acquisition from these experts with and without access to machine learning facilities. The use of machine learning with knowledge acquisition from experts led to the production of significantly more accurate rules in significantly less time than knowledge acquisition from experts alone. This combination also led to more accurate rules than the use of machine learning alone, a difference that was statistically significant for the Glass domain.

At the very least, this study has demonstrated that there are contexts in which the integration of machine learning with knowledge acquisition from experts is beneficial. It remains for future research to map in detail the types of knowledge acquisition task for which such integration is advantageous. Although the current study has demonstrated benefit in one context, it provides little evidence about the range of contexts for which this will occur or whether alternative approaches to the integration will give similar results. Ideally, such studies would be conducted with genuine domain experts engaged in real-world knowledge acquisition tasks, but such large-scale experimental studies imply tremendous cost. In consequence, evidence may have to be gathered by the less comprehensive but much more feasible combination of circumscribed experimental studies, such as the current one, together with individual case studies of real-world applications.

The current study has focused on a restricted part of the knowledge acquisition cycle—the formulation, testing, and refinement of rules once an appropriate class of model and vocabulary have been defined. It has demonstrated that The Knowledge Factory's integration of machine learning with knowledge acquisition from experts can offer benefits at this stage of the cycle. It would be straightforward to conduct similar studies using other knowledge acquisition software for the same tasks, which would help identify the specific features of our approach that conferred benefit in this context. It would also be valuable to conduct similar studies on different knowledge acquisition tasks in order to delimit the types of task for which these techniques are beneficial. A more ambitious extension to this type of study would examine larger scale tasks that included the formulation of appropriate ontologies. It would be much more difficult to perform controlled experiments in such a context, however, as it is not clear how to supply examples to subjects without strongly influencing the ontology that is selected, because an ontology is required to describe the examples. The study of tasks that involve non-trivial selections between classes

of model, such as a choice between attribute-value or first-order representations, may be even more difficult, as most tools support only very limited variations in the type of model employed.

9. Conclusions

Integration of machine learning with knowledge acquisition from experts has considerable intuitive appeal. These two approaches to knowledge acquisition have different features that appear to complement one another. In consequence, many techniques have been developed for integrating them. However, there has been little formal evaluation of the effectiveness of these integrated techniques.

The current study has demonstrated that integration of machine learning with knowledge acquisition can increase the accuracy of the knowledge bases developed. The expert systems created in this study through integrated use of both methods were more accurate than those developed by either technique in isolation. This increase in accuracy was accompanied by a decrease in acquisition time in comparison to knowledge acquisition without machine learning. This second outcome suggests that the integration of machine learning can make knowledge acquisition less difficult. This conclusion is reinforced by subjects' subjective judgments. A further benefit of the integrated approach is that it can increase the developer's confidence in the accuracy of the resulting expert system. Finally, questionnaire results indicated a very positive response to the manner in which machine learning was integrated into The Knowledge Factory software.

There are a wide variety of techniques for integrating machine learning with knowledge acquisition from experts. Those examined in this study are distinguished by being oriented for direct use by domain experts with little knowledge engineering expertise. As the experiment employed subjects of this type, it provides support for the efficacy of these techniques in this context, although there is need for further research to delineate the scope of the approach.

Acknowledgments

This research has been supported by the Australian Research Council and the Apple University Development Fund. We are grateful to Tim Menzies for encouraging us to use undergraduate students as experimental subjects for knowledge acquisition research. This paper has benefited greatly from the reviewers' suggestions and Pat Langley's editorial comments.

The study presented herein replicates aspects of an earlier study (Webb & Wells, 1996), but rectifies problems with the previous experiment, including a serious deficiency in the user interface of the software. The earlier paper discusses these problems in detail.

Notes

1. *Machine learning* encompasses a wide variety of automated processes including explanation-based learning and conceptual clustering. This paper concerns what Langley (1996) calls *rule learning*, that is, systems for inferring logical models from data, such as decision tree or classification rule learners. For ease of exposition, we use the term *machine learning* throughout the paper in this restricted sense.
2. One subject did not answer questions 5 and 6 from which 3a and 4a were derived and hence was excluded from this and the next analysis.

References

Attar Software (1989). *Structured decision tasks methodology for developing and integrating knowledge base systems.* Attar Software, Leigh, Lancashire.

Boose, J. H. (1986). ETS: A system for the transfer of human expertise. In J. S. Kowalik (Ed.), *Knowledge based problem solving.* New York: Prentice-Hall.

Buntine, W., & Stirling, D. (1991). Interactive induction. In J. E. Hayes, D. Michie, & Ė. Tyugu (Eds.), *Machine Intelligence 12.* Oxford: Clarendon Press.

Compton, P., Edwards, G., Srinivasan, A., Malor, R., Preston, P., Kang, B., & Lazarus, L. (1992). Ripple down rules: Turning knowledge acquisition into knowledge maintenance. *Artificial Intelligence in Medicine*, *4*, 47–59.

Davis, R., & Lenat, D. B. (1982). *Knowledge-based systems in artificial intelligence.* New York: McGraw-Hill.

De Raedt, L. (1992). *Interactive theory revision.* London: Academic Press.

Gams, M., Drobnič, M., & Karba, N. (1996). Average-case improvements when integrating ML and KA. *Applied Intelligence*, *6*, 87–99.

Kodratoff, Y., & Vrain, C. (1993). Acquiring first order knowledge about air traffic control. *Knowledge Acquisition*, *5*, 1–36.

Langley, P. (1996). *Elements of Machine Learning.* San Francisco: Morgan Kaufmann.

Le Grand, A., & Sallantin, J. (1994). A framework to improve knowledge acquisition based on machine learning. *Proceedings of the Eleventh European Conference on Artificial Intelligence* (pp. 493–497). London: John Wiley.

Merz, C. J., & Murphy, P. M. (1997). UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine.

Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* Berlin: Springer-Verlag.

Morik, K. (1987). Aquiring domain models. *International Journal of Man-Machine Studies*, *26*, 93–104.

Morik, K., Wrobel, S., Kietz, J.-U., & Emde, W. (1993). *Knowledge acquisition and machine learning: Theory, methods, and applications.* London: Academic Press.

Nedellec, C., & Causse, K. (1992). Knowledge refinement using knowledge acquisition and machine learning methods. *Proceedings of the 1992 European Knowledge Acquisition Workshop* (pp. 171–190). Berlin: Springer-Verlag.

Nedellec, C., Correia, J., Ferreira, J. L., & Costa, E. (1994). Machine learning goes to the bank. *Applied Artificial Intelligence*, *8*, 593–615.

O'Neil, J. L., & Pearson, R. A. (1987). A development environment for inductive learning systems. *Proceedings of the 1987 Australian Joint Artificial Intelligence Conference* (pp. 673–680). Sydney.

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

Schmalhofer, F., & Tschaitschian, B. (1995). Cooperative knowledge evolution for complex domains. In G. Tecuci & Y. Kodratoff (Eds.), *Machine learning and knowledge acquisition: Integrated approaches.* London: Academic Press.

Smith, R. G., Winston, H. A., Mitchell, T. M., & Buchanan, B. G. (1985). Representation and use of explicit justifications for knowledge base refinement. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 673–680). San Mateo, Ca: Morgan Kaufmann.

Sommer, E., Morik, K., André, J.-M., & Uszynski, M. (1994). What online machine learning can do for knowledge acquisition—A case study. *Knowledge Acquisition*, *6*, 435–460.

Tecuci, G., & Kodratoff, Y. (1990). Apprenticship learning in imperfect domain theories. In Y. Kodratoff & R. Michalski (Eds.), *Machine learning: An artificial intelligence approach.* San Mateo, CA: Morgan Kaufmann.

Tecuci, G. (1995). Building knowledge bases through multistrategy learning and knowledge acquisition. In G. Tecuci & Y. Kodratoff (Eds.), *Machine learning and knowledge acquisition: Integrated approaches.* London: Academic Press.

Webb, G. I. (1992). Man-machine collaboration for knowledge acquisition. *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence* (pp. 329–334). Singapore: World Scientific.

Webb, G. I. (1993). DLGref2: Techniques for inductive knowledge refinement. *Proceedings of the IJCAI Workshop on Machine Learning and Knowledge Acquisition: Common Issues, Contrasting Methods, and Integrated Approaches* (pp. 236–252). Chambery, France.

Webb, G. I. (1996). Integrating machine learning with knowledge acquisition through direct interaction with domain experts. *Knowledge-Based Systems*, *9*, 253–266.

Webb, G. I., & Agar, J. W. M. (1992). Inducing diagnostic rules for glomerular disease with the DLG machine learning algorithm. *Artificial Intelligence in Medicine*, *4*, 3–14.

Webb, G. I., & Wells, J. (1995). Recent progress in machine-expert collaboration for knowledge acquisition. *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence* (pp. 291–298). Singapore: World Scientific.

Webb, G. I., & Wells, J. (1996). Experimental evaluation of integrating machine learning with knowledge acquisition through direct interaction with domain experts. *Proceedings of the Pacific Knowledge Acquisition Workshop* (pp. 170–189). Sydney, Australia.

Wilkins, D. C. (1988). Knowledge base refinement using apprenticeship learning techniques. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 646–651). San Mateo, CA: Morgan Kaufmann.