P92-36

A Method for Evaluation of Learning Systems

Tomasz Arciszewski, Tomasz Dybala, and Janusz Wnek*

ABSTRACT—The development of a method for evaluating the performance of learning systems is a prerequisite to their engineering applications. Such a method is also necessary for machine learning researchers for the objective evaluation and comparison of the performance of individual experimental and commercial learning systems. This paper proposes such a method, based on the performance of a learning system in the entire multistage process of automated knowledge acquisition. The evaluation utilizes both the values of individual criteria selected for a given evaluation and the character of their learning curves. The description of the method includes its basic assumptions, the evaluation process, evaluation criteria and their classification, and an evaluation model based on the multi-attribute utility theory. Individual assumptions are justified, taking into consideration the results produced using the AQ15 learning system on two collections of actual engineering examples from the areas of construction safety and structural design. Recommendations for further research and conclusions are also provided.

Keywords: Learning engineering, learning system, performance evaluation, evaluation criteria, evaluation model, multi-attribute utility theory.

Introduction

In recent years, research on machine learning has resulted in the development of many experimental and commercial learning systems such as pLogic [9], DQuest [10], DataLogic [11], SuperExpert [13], BEAGLE [14], Aurora [4], and INLEN [6]. However, these systems are still being used mostly for research purposes, and their engineering applications are limited [7, 12]. This delay can be explained by the methodological gap between the machine learning and engineering communities. Currently, computer scientists are concerned with the internal workings of learning systems and have only limited interest in the methodology of using these systems in the process of knowledge acquisition. For engineers, learning systems could be useful tools. However, the lack of methodologies for their use delays practical applications. Therefore, *learning engineering*, a new subarea of knowledge engineering which will provide methodologies of automated knowledge acquisition and will fill the presently existing gap between machine learning and engineering, should be developed. A particularly important part of learning engineering is a method for the evaluation of learning systems. Its development is a prerequisite to the practical application of learning systems. This method is also necessary for machine learning researchers, for the objective evaluation and comparison of learning systems.

There have been attempts to evaluate various learning systems and to generalize this experience in the form of evaluation criteria. Weiss and Kulikowski [15] investigated the problem of statistical estimation of the true performance of a learning system in classification tests. This true performance is measured by the true error rate for classification tests, and it is statistically defined as the error rate of the learning system on an asymptotically large number of new examples converging to the actual population distribution. Since the number of available examples is usually limited, the true error rate can only be estimated. This can be done using estimators, empirical error rates, whose values are determined utilizing different resampling methods. The methods discussed by Weiss and Kulikowski [15] include the holdout method and two resampling methods: leave-one-out and ten-fold cross-validation. These methods have also been adopted in the evaluation method proposed in this paper.

In another project, Wnek and Michalski [16] experimentally compared five different systems for concept learning from examples. The systems represented different learning paradigms, i.e., symbolic learning: the decision tree learning system—C4.5, the rule learning system—AQ15, the rule learning with constructive induction system—AQ17-HCI, the backpropagation

*Tomasz Arciszewski is on sabbatical leave from Intelligent Computers Laboratory, Civil Engineering Department, Wayne State University, Detroit, MI. Tomasz Dybala and Janusz Wnek can be contacted at the Center for Artificial Intelligence, Computer Science Department, George Mason University, Fairfax, VA 22030

HIEURISTICS, The Journal of Knowledge Engineering, Special Issue on Knowledge Acquisition and Machine Learning, Vol. 5, No. 4, Winter 1992.

system— BpNet, and a genetic algorithm based classifier system—CFS. The systems were tested on logic-style concepts (in disjunctive normal form) from the artificial domain. The use of a well-defined domain allowed for a comparison in terms of the exact error rate and learned description complexity. For the first time, a diagrammatic visualization technique was used to display an error image. The exact error rate was defined as the ratio between exact error (the cardinality of the set-difference between the union and the intersection of the target and learned concepts) and the size of the instance space.

In the Monks project [16], sixteen learning systems were compared, considering the true error rate for a small but complete population of examples. This error rate was determined using the holdout method for various ratios of learning to testing examples. The comparison study was conducted by twelve research teams from several countries.

In the projects reported, the evaluation of performance of learning systems was based on their performance in a one-stage automated knowledge acquisition process conducted for a given number of examples. Only the final error rates, calculated for all examples available, were determined. This evaluation did not cover the dynamics of the learning process in terms of changes in decision rules (and their sensitivity to new examples added to the learning examples) and error rate changes which occur over individual stages of the automated knowledge acquisition process. The approach used can be compared to the evaluation of the behavior of a complex system by considering its performance only at a given stage of a multistage process, when instead the entire process and its dynamics and sensitivity should be considered. Therefore, the evaluation of a learning system should be based on the system's behavior in a multistage automated knowledge acquisition process, which is described by a class of learning curves for individual evaluation criteria. Only such an evaluation will provide a global understanding of a given learning system and enabling it to be compared with other systems, producing results meaningful for both engineers and machine learning researchers.

This paper reports the results of research on the evaluation of learning systems conducted at the Artificial Intelligence Center at George Mason University in Fairfax, Virginia. The objective of this research was to develop a universal evaluation method utilizing various evaluation criteria. The objective of this paper is to propose a general outline of the evaluation method and to discuss a group of evaluation criteria known as error rates. The method is intended for both engineers and machine learning researchers for evaluating and comparing learning systems. The paper includes a discussion of the assumptions made by the proposed method, its procedure, performance-based evaluation criteria and their classification, and an evaluation model based on multi-attribute utility theory. Assumptions are explained with reference to results produced using the AQ15 learning system as applied to two collections of actual engineering examples from the areas of construction safety and structural design. Recommendations for the further research and conclusions are also given.

Evaluation Method

Basic Assumptions

The proposed method is based on the following assumptions:

1. Learning is rule generation.

2. The automated knowledge acquisition process is a multistage process of knowledge acquisition in which the learning system is used as a knowledge acquisition tool.

3. The evaluation of a learning system is understood as the evaluation of its performance in the automated multistage knowledge acquisition process. Also, it reflects its behavior during this process. The behavior is described by a class of learning curves for individual evaluation criteria.

4. Evaluation is based on the results of knowledge acquisition conducted for one or more collections of examples. If more than one collection is used, they should include both 'noisy' and 'clean' collections in terms of errors in classification and in the values of individual attributes.

5. For each collection of examples, the knowledge acquisition process should be conducted for several different randomly generated sequences of examples.

6. The evaluation of performance is conducted using all or selected evaluation criteria. These criteria are considered to be estimators of the true performance of the learning system.

7. The final evaluation is conducted using a multi-attribute utility evaluation model.

Evaluation Procedure

The evaluation procedure proposed is general and can be

used for any evaluation criteria.

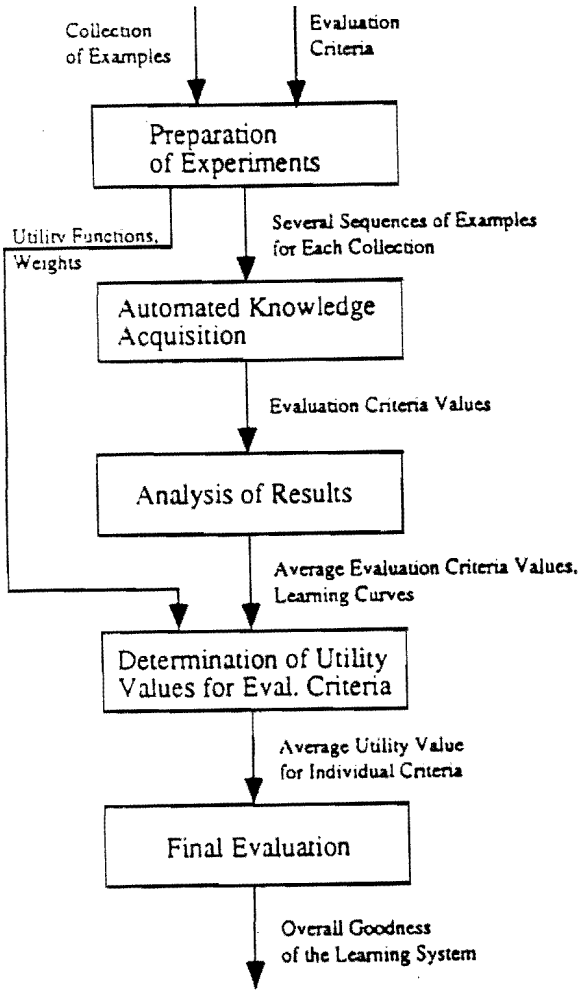

Figure 1. Evaluation Procedure

It has five major phases:

1. Preparation of experiments.
2. Automated knowledge acquisition.
3. Analysis of results.
4. Determination of utility values for evaluation criteria.
5. Final evaluation.

The first phase, Preparation of Experiments, has the following six stages:

1. Examination of examples.
2. Numbering of examples.
3. Division of examples.
4. Final grouping of examples.
5. Determination of importance of a given evaluation criterion.
6. Preparation of utility functions.

In the Examination of Examples stage, all examples are examined to determine if they are complete and correct, and if their form is compatible with the learning system being tested. Incomplete, incorrect, and incompatible examples are removed. In the second stage, examples are numbered. In the third stage, random sequences of numbers are generated for each collection of examples. Each sequence of numbers represents a sequence of examples used for individual automated knowledge acquisition processes. In the fourth stage, examples are divided into several groups. The number of groups is equal to the assumed number of stages in the automated knowledge acquisition process. Group 1 is taken for the first stage; groups 1 and 2 are added for the second stage; groups 1, 2, and 3 are added for the third stage; etc. In the fifth stage, the relative importance of individual evaluation criteria, represented by weights, is introduced. In the last stage, single attribute utility functions are prepared using ranking, rating, or bisection techniques. A computer package, RAMZES, based on these techniques, can also be utilized [1, 2].

In the second phase, Automated Knowledge Acquisition, knowledge acquisition processes are conducted for all collections of examples. In each case, the learning system produces decision rules and conducts tests. In these tests, the generated rules are matched against testing examples in accordance with the requirements of individual testing methods, as described in a later section.

In the third phase, Analysis of Results, the results of the automated knowledge acquisition are used to determine the values of evaluation criteria for all collections of examples and all their sequences. For all collections of examples, the average evaluation criteria values are calculated over all the sequences of examples considered, and these values are used to construct learning curves for average evaluation criteria. A learning curve is understood here as a relationship between a given evaluation criterion and the number of examples used to produce its value. Learning curves are determined using regression analysis on the values of individual evaluation criteria.

In the fourth phase, Determination of the Utility Value, the following stages take place for all collections of examples:

1. Classification of learning curves.
2. Determination of modified average evaluation criteria values.
3. Determination of utility values of modified average

evaluation criteria values.

4. Calculation of average utility values for individual evaluation criteria.

In the first stage of this phase, the shapes of the learning curves for average evaluation criteria for all collections of examples are analyzed, and these curves are classified as either desirable, acceptable, or undesirable. The level of desirability is measured by the modification factor $m \in [0.5 .. 1.0]$. The classification is based on the experience of the human expert. Obviously, this is somewhat subjective in character, but it is important to incorporate into the evaluation process the behavior of the learning system in the entire automated knowledge acquisition process. It is proposed to assign modification factor values 1.0, 0.75 and 0.5 for desirable, acceptable, and undesirable curves respectively. In the second stage, average evaluation criteria values, which were obtained for all collections of examples, are multiplied by the modification factors determined for these criteria, and thus modified average evaluation criteria values are determined. In the next stage, the single-attribute utility functions, identified in phase one of the evaluation procedure, are used to determine the utility values of individual modified average evaluation criteria values. These values are used next in the last stage of this phase to calculate average utility values for the individual evaluation criteria.

The last phase of the evaluation process, Final Evaluation, is initiated when values of utilities are known for all evaluation criteria. The final evaluation is done using the evaluation model: all average utility values for evaluation criteria are multiplied by their weights and added to produce the final value of the utility of the learning system. This value represents the measure of the overall "goodness" of the learning system evaluated.

Multi-Attribute Utility Evaluation Model

It is assumed that a given learning system, LS, is described by a set $EC = \{EC_i\}$, $i = 1,...,n$ of n evaluation criteria. For each criterion EC_i, a conditional utility function on EC_i can be determined, such that this function maps EC_i onto [0,1]. It has also been assumed that for each evaluation criterion, preferences of levels for this criterion are not affected by the fixed levels of any other criterion and therefore the mutual independence of preferences holds [5]. Under these assumptions, the additive integration rule can be used to determine the utility value u(LS) (a measure of goodness or worth) for a given learning system.

$$u(LS) = \sum_{i=1}^{n} w_i u(EC_i)$$

where: LS - learning system under evaluation

u(LS) - utility value of the system LS

w_i - weight of the ith evaluation criterion EC_i

$$\sum_{i=1}^{n} w_i = 1$$

Evaluation Criteria

The performance of a learning system can be evaluated using various evaluation criteria. However, error rates are considered the most useful and meaningful for practical purposes in science and engineering [15] and were initially selected to conduct evaluation experiments and to evaluate the proposed method.

A class of evaluation criteria is proposed. It has been developed to meet the needs of both engineers and machine learning researchers. All proposed criteria can be used in a given evaluation singly or in combination, based on the needs and preferences in a given evaluation.

When error rates are considered, three major cases of available data can be distinguished:

1. A large population of examples is available, and the system can be evaluated on an asymptotically large number of examples converging to the actual population distribution. In this case *the true error rates* can be determined.

2. A small but complete population of examples is available, enabling the system to be evaluated using all these examples. In this case, *the true error rates for a small population* can be determined.

3. Only a sample of examples from a large population is available, and the system can be evaluated using only these examples. In this case, *the empirical error rates* can be determined, which can be used as estimators of the true error rates.

In knowledge acquisition in science and engineering, the third case is the usual one and therefore it is the most important from the practical point of view. For this reason, we decided to consider only the empirical error rates in this paper.

Four major classes of empirical error rates are distinguished:

1. Overall empirical error rates.
2. Omission empirical error rates.

3. Commission empirical error rates.
4. Combined empirical error rates.

Overall empirical error rates are defined as:

$$E_{ov} = \frac{\text{number of errors}}{\text{number of tests}}$$

where:

an error is a misclassification of a testing example and number of tests is the number of classification tests

Omission empirical error rates are defined as:

$$E_{om} = \frac{\sum_{i=1}^{n} E_{om}^{i}}{n}$$

where:

n - the number of classes

$$E_{om}^{i} = \frac{\text{number of omission errors for class "i"}}{\text{number of tested positive examples of class "i"}}$$

number of omission errors for class "i" - number of errors when a positive example is classified as a negative one

number of tested positive examples for class "i" - number of classification tests for class "i" examples

Commission empirical error rates are defined as:

$$E_{co} = \frac{\sum_{i=1}^{n} E_{co}^{i}}{n}$$

n - the number of classes

$$E_{co}^{i} = \frac{\text{number of comission errors for class "i"}}{\text{number of tested negative examples of class "i"}}$$

number of commission errors for class "i" - number of errors when a negative example is classified as a positive one

number of tested negative examples for "i" class - number of classification test for negative "i" class examples

Combined empirical error rates are defined as:

$E_{cb} = \alpha E_{om} + (1-\alpha) E_{co}$, where:

α - importance factor, $\alpha \in [0..1]$

An identical classification of error rates for all four classes of empirical error rates is proposed. It is based on five classification attributes which are related to the most important methodological aspects of the calculated empirical error rates. These attributes, the relevant question, and possible answers are given below. Names for evaluation criteria identified by individual values of attributes are also given. Each empirical error rate can then be characterized (identified) as a combination of the values of these five attributes. Attributes are given in the same sequence as in decision making regarding the selection of evaluation criteria.

1. Number of Collections
   How many collections of examples are used?
   1.1. One - single collection error rate
   1.2. More than one - multi-collection error rate or average error rate

2. Number of sequences
   How many sequences of examples are used?
   2.1. One - single sequence error rate
   2.2. More than one - multi-sequence error rate or average error rate

3. Division of Examples
   Are examples divided into training and testing examples?
   3.1. No - apparent error rate
   3.2. Yes - empirical error rate

4. Division Type
   How many times is the division of examples into training and testing sets conducted?
   4.1. No division - apparent error rate
   4.2. Single division - holdout error rate
   4.3. Multi-division - cross validation error rate

5. Resampling Method
   How is the multi-division (resampling) conducted?
   5.1. No resampling - apparent or holdout error rate
   5.2. One example is used as testing example - leave-one-out cross validation error rate
   5.3. "k" examples are used as testing examples - k-fold cross validation error rate

The classification attributes for all classes of error rates are given in Table 1. In this table, individual rows represent subsequent attributes and their possible values.

When attributes "number of collections" and "number of sequences" are considered, the most important decision is to analyze one collection/sequence or more than one collection/sequence. When more than one collection/sequence is analyzed, error rates for individual collections/sequences can be averaged, and this provides

No	Attribute Name	Possible Values		
		1	2	3
1	Number of collections	One	More than one	
2	Number of Sequences	One	More than one	
3	Division of examples	No	Yes	
4	Division type	No division	Single division	Multi division
5	Resampling method	No resampling	No resampling	K-fold cross valid

Table 1. Empirical Error Rates: Classification Attributes and Their Values

a more generalized performance evaluation than the one based on a single collection/sequence. The attribute "division of examples" describes how the available examples are divided into training and testing examples. When no division is conducted and the same examples are used for training and testing, error rates determined on such examples are called "apparent." When a division is conducted, it is necessary to decide how to divide examples. The attribute "division type" addresses this issue. When examples are divided many times in the process of resampling, the attribute's value is "multi-division." When this division is conducted only once, as in the holdout method, the attribute's value is "single division."

The attribute "resampling method" determines how the learning examples are selected. Three basic cases are distinguished. In the first case, no resampling is conducted and the division of examples is conducted only once, in this case the holdout method is utilized. When resampling is performed, examples are divided many times into separate groups of learning and testing examples, and testing is conducted on examples which were not used for learning. Two cases are distinguished, according to the method of subsampling used: leave-one-out cross-validation, or k-fold cross-validation [15]. In the leave-one-out method, tests are conducted on all examples. For each test, one example is removed from the collection of available examples, and the remaining examples are used for learning. The decision rules produced are then used to classify the examples. Learning is conducted separately for each test and for a different combination of examples. In the k-fold method, all available examples are randomly divided into training and testing examples, k percent of examples is used for testing, and the remaining (100 minus k) percent for learning. For a given division of examples, learning is conducted only once on training examples, but the entire procedure is repeated k times and the final error rate is determined as an average error rate considering the error rates obtained for individual combinations of training and testing examples. In the holdout method mentioned earlier, two-thirds of the examples are used for learning and the remaining one-third for testing. The test is conducted only once. The classification tree, identical for all four classes of empirical error rates, is shown in Figure 2. For simplicity, only one branch is presented, because the second branch is identical.

Justification of the Method's Assumptions

In an earlier section, four classes of error rates were proposed. The use of more than one error rate in a given evaluation is recommended, because the learning curves for individual error rates usually differ, and a single learning curve might not provide the desired understanding of the learning system's behavior in the automated knowledge acquisition process. For example, in Figure 3, learning curves for the overall, omission, and commission error rates are shown. These error rates were determined using the holdout method of example division and resampling for the collection of 336 examples of computer-generated minimum weight designs of wind bracings in tall buildings [8].

It is important to evaluate a given learning system on more than one collection of examples, because learning

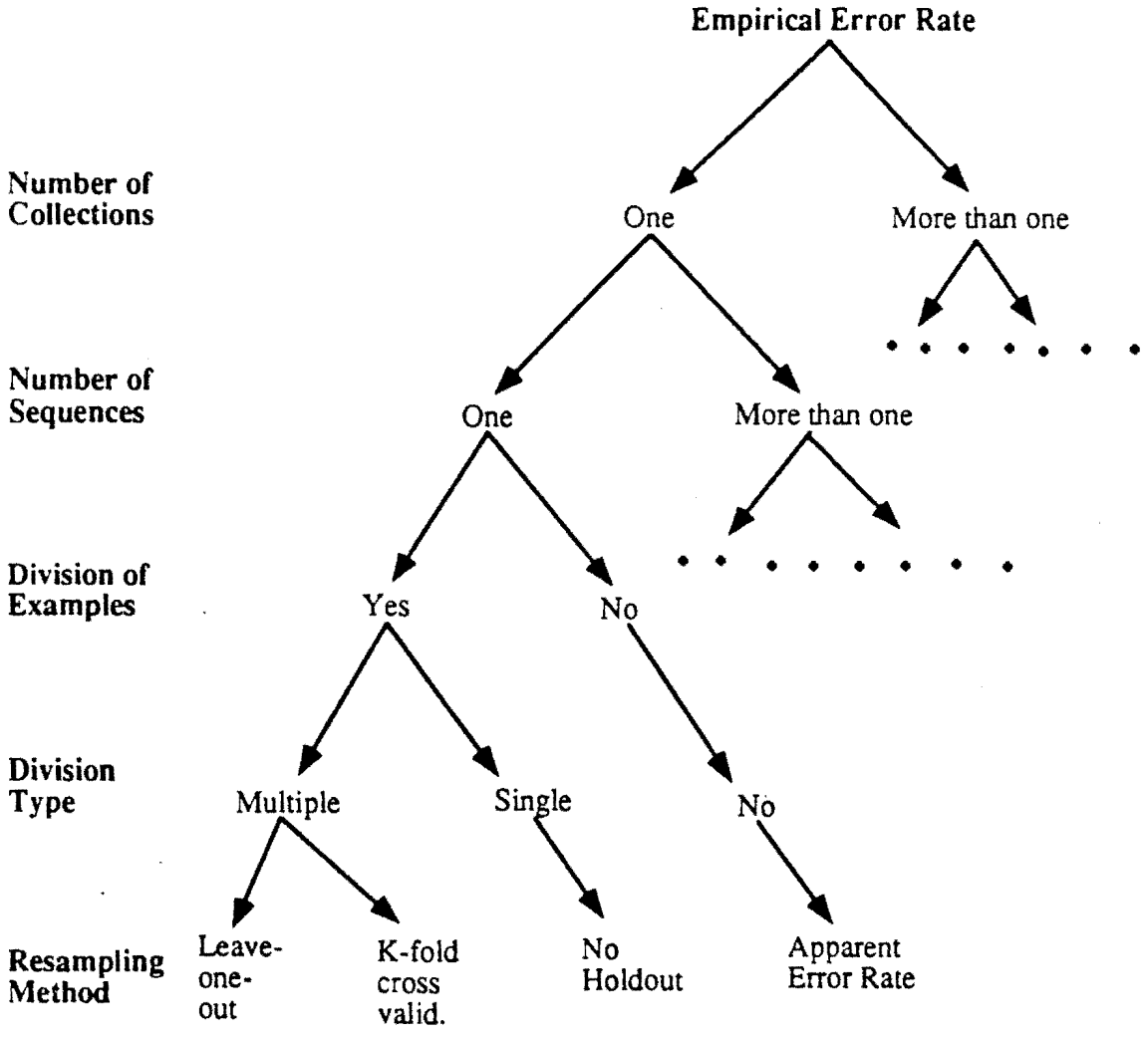

Figure 2. Classification of Empirical Error Rates

systems usually display different performance on "noisy" and "clean" data. It has been observed by the authors that some systems perform better than others on "noisy" data and significantly worse than others on "clean" data. For this reason, it is prudent to base the evaluation on both types of data. To illustrate the variation in performance on different data, Figure 4 shows two learning curves for average overall error rates. One curve was obtained for a collection of 225 examples of construction accidents [3], while the other was for a collection of 336 examples of

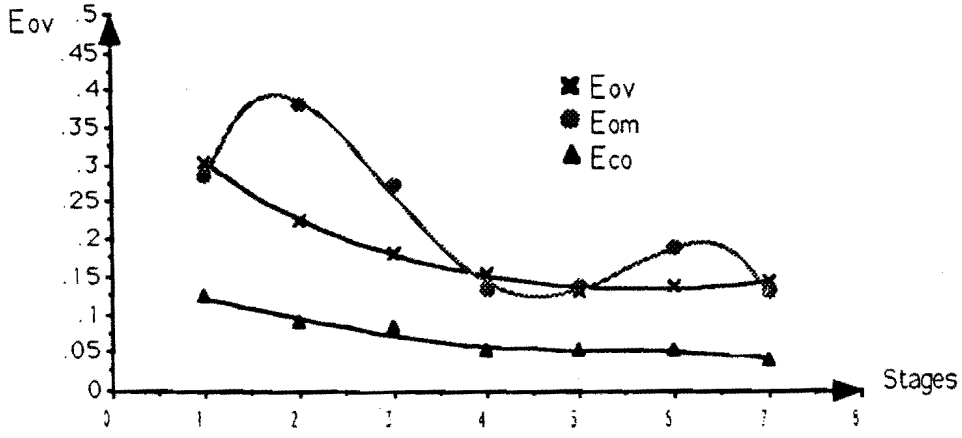

Figure 3. Comparison of Learning Curves for Overall, Omission and Commission Error Rates

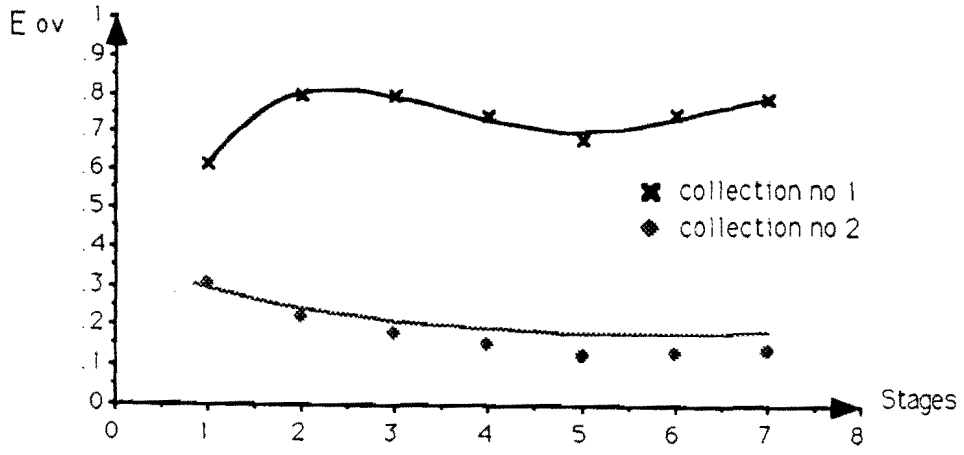

Figure 4. Comparison of Learning Curves for the Overall Error Rate for Two Different Collections of Examples

computer-generated minimum weight designs of wind bracings in tall buildings [8].

How the division of examples and their resampling is conducted also strongly affects the final error rates and the shapes of the learning curves. Figure 5 shows three learning curves obtained for overall error rates determined for the same collection of structured examples [8] using the holdout method and the leave-one-out and ten-fold cross validation methods. For this reason, all three methods of division of examples and their resampling should be used to produce a complete picture of the behavior of a given learning system in the automated knowledge acquisition process.

It is also important to conduct knowledge acquisition on several sequences of examples from each collection, to minimize the effects of random selection of examples for individual stages in the automated knowledge acquisition process. Figure 6 shows learning curves for overall empirical error rates for individual sequences of examples for the same collection of examples as in Figure 5. There are significant differences in the shapes of these curves for various sequences, and results obtained only for a single sequence of examples might be misleading in terms of the nature of the automated knowledge acquisition process.

Conclusions

The evaluation of learning systems is an important area of learning engineering. The progress in this area will have a significant impact on the applicability of machine learning to engineering. The development of a method for evaluation and comparison of learning systems is considered a prerequisite to practical application, and for this reason research on the evaluation of learning systems was initiated at the Artificial Intelligence Center at George Mason University. The methodological and experimental results presented in this paper were produced in the initial stage of this research. However, the evaluation method

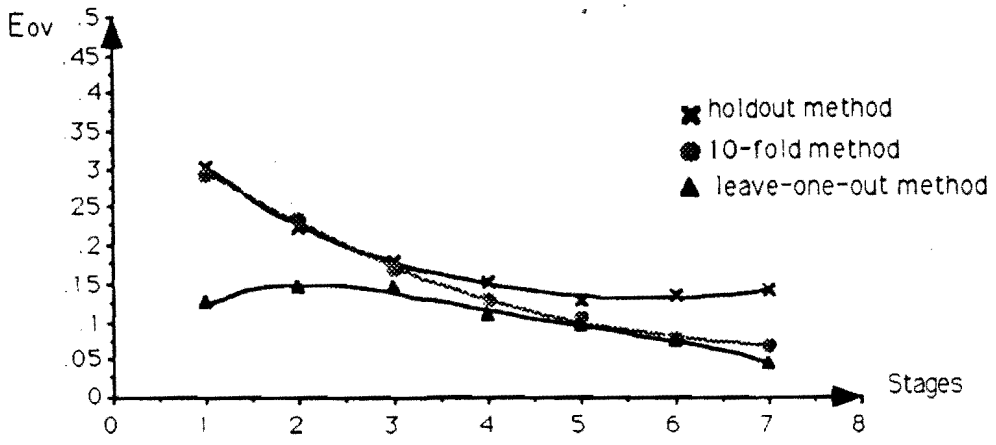

Figure 5. Comparison of Learning Curves for the Average Overall Error Rate for the Holdout, the Leave-one-out, and the Ten-Fold Cross Validation Methods

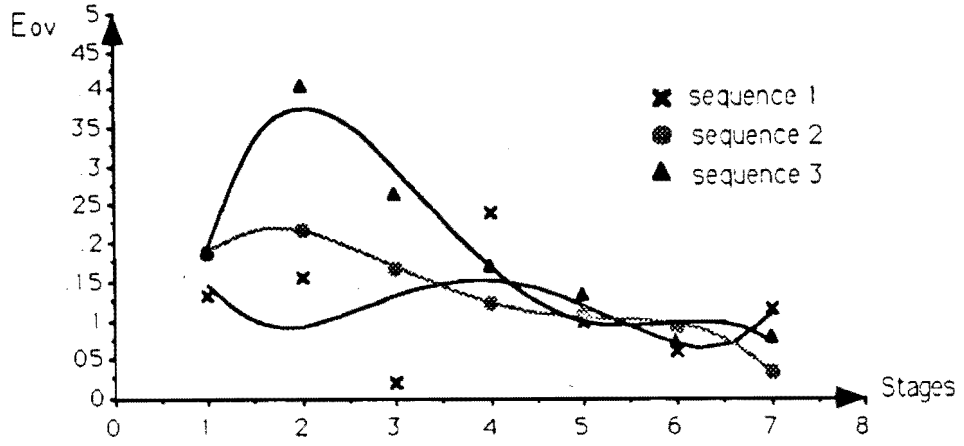

Figure 6. Comparison of Learning Curves for the Average Overall Error Rate for Three Different Sequences of Examples

proposed is methodologically complete and can be used with any evaluation criteria.

The ultimate objective of the evaluation described here is to develop a good understanding of the behavior of a learning system in the automated knowledge acquisition process. This can be accomplished using the proposed method and various evaluation criteria, not necessarily performance-based, also proposed in this paper. However, empirical error rates provide important quantitative evaluation measures which are convincing for both machine learning researchers and engineers. For this reason, these criteria are proposed for current use, before more adequate criteria are developed and their feasibility proven.

The evaluation method proposed here is based on several basic assumptions, which are justified by appropriate experimental results. The experiments demonstrated significant differences in performance by a learning system for individual evaluation criteria. Therefore, the application of several criteria in a given evaluation is recommended to produce results more meaningful than those which could be obtained for a single criterion. The performance of a learning system depends strongly on the noise level in the examples, and therefore more than one example collection should be used in an evaluation, to reduce the effect of this noise. Preferably, both "noisy" and "clean" collections of examples should be used. It has been observed that the division of examples and their resampling method affect the empirical error rates. For this reason, all three methods of division of examples and their resampling should be used. The sequencing of examples also has an impact on the empirical error rates, and to minimize it, several sequences should be used and the results averaged.

Future research on the evaluation of learning systems should result in the development of a standard evaluation method which would be acceptable to both machine learning experts and potential users of learning systems. Also, several typical collections of actual engineering examples with different levels of noise should be prepared as bench-marks. Future research should also address the issue of development and verification of various evaluation criteria, particularly those which are not performance-based. The evaluation of learning systems is a complex area, dealing with the development of evaluation procedures, evaluation criteria, and machine learning itself. The development of a standard evaluation method is a challenge, but one which must be met if practical applications of machine learning are to take place.

Acknowledgements

The authors would like to express their gratitude to Ryszard S. Michalski for his helpful suggestions and comments. We also acknowledge the contribution of Muntaz Usmen, Civil Engineering Department, Wayne State University, who prepared examples of construction accidents, and of Mohamad Mustafa of the same department, who prepared the examples of structural wind bracings designs.

This research was done in the GMU Center for Artificial Intelligence. Research at the Center is supported in part by the Defense Advanced Research Projects Agency under grants administered by the Office of Naval Research (No. N00014-91-J-1854), and grant No. F49620-92-J-0549 administered by the Air Force Office of Scientific Research, in part by the Office of Naval Research under grant (No. N00014-91-J-1351), and in part by the National Science Foundation Grant (No. IRI-9020266).

References

1. Arafat, G. 1991. "Concept Evaluation in Structural

Design Utilizing Multi-attribute Utility Theory and a Knowledge-Based Approach," *Ph.D. Thesis*, Wayne State University, Detroit, Michigan.

2. Arafat. G., Goodman, B. and Arciszewski, T. 1991. "RAMZES: A Knowledge-Based System for Structural Concept Evaluation," *Proceedings. of the Second Intl. Conf. on the Applications of AI Techniques to Civil and Structural Engineering*, Oxford, England.

3. Arciszewski, T., Usmen, M. and Gleichman, J. 1991. "Construction Accident Analysis: The Inductive Learning Approach," *Proceedings. of the ASCE Congress on Construction*, Boston, MA.

4. INIS Inc. 1988. "User's Guide to AURORA 2.0. A Discovery System," Fairfax, VA.

5. Keeney, R. and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York, NY.

6. Michalski, R.S., Kerschberg, L., Kaufman, K. and Ribeiro, J. 1992. "Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation and First Results," *International Journal on Intelligent Systems*, No. 1.

7. Modesitt, K. 1987. "Space Shuttle Main Engine Anomaly Data and Inductive Knowledge-based Systems: Automated Corporate Expertise," *Proceedings of the Conference on Artificial Intelligence in Space Research*, Huntsville, Alabama.

8. Mustafa, M. and Arciszewski, T. 1992. "Inductive Learning of Wind Bracings Design," in *Knowledge Acquisition in Civil Engineering*, T. Arciszewski.and L. Rossman (Eds.), ASCE.

9. pLOGIC Inc. 1991. "pLogic: Reference Guide," Torrance, CA.

10. REDUCT SYSTEMS Inc. 1991. "DataQuest," Regina, Canada.

11. REDUCT SYSTEMS Inc. 1992. "DataLogic," Regina, Canada.

12. Sillen, R.V. 1987. "Building PC-hosted Expert Systems using Induction Methods," *Research Report, Novocast Expert Systems A.B*, Sweden.

13. Softsync.Inc. 1986. "Super Expert," New York.

14. Warm Boot Ltd. 1987. "PC/BEAGLE: User Guide," Nottingham, England.

15. Weiss, S.M. and Kulikowski, C.A. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*, Morgan Kaufmann Publishers.

16. Wnek, J. and Michalski, R.S. 1992. "Experimental Comparison of Symbolic and Subsymbolic Learning" *Heuristics, Special Issue on Knowledge Acquisition and Machine Learning.*