Machine Learning, 18, 109–114 (1995)

Research Note

Classification Accuracy: Machine Learning vs. Explicit Knowledge Acquisition

ARIE BEN-DAVID AND JANICE MANDEL msariebd@pluto.mscc.huji.ac.il
Management Information Systems, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905 Israel

Editor: Bruce Porter

Abstract. This empirical study provides evidence that machine learning models can provide better classification accuracy than explicit knowledge acquisition techniques. The findings suggest that the main contribution of machine learning to expert systems is not just cost reduction, but rather the provision of tools for the development of better expert systems.

Keywords: Expert systems, machine learning, explicit vs. implicit knowledge acquisition, classification accuracy

1. Introduction

Can machine learning offer anything to expert systems? B.G. Buchanan (1989) raised this important question in an interesting article under this title in *Machine Learning*. "The commercial world of expert systems at large seems unconvinced that machine learning has anything to offer yet. I strongly disagree," he wrote five years ago. Despite of the progress which has been made since, Buchanan's observations are still quite valid. Perhaps the main reason for the skeptic attitude toward machine learning (ML) in commercial circles stems from the fact that the academic world has not presented yet sufficient evidence which justifies abandoning older, relatively well established, methods in favor of ML models. Surprisingly many firms are involved now in expert systems (Ansari & Modarress, 1990). How can one convince a company which currently develops or uses expert systems (ES) that ML can be of substantial benefit? Developers of expert systems mainly use *explicit knowledge acquisition* (EKA). EKA is known as very costly and time consuming, mainly since experts frequently find it difficult to explain how they make their decisions (Hays-Roth, 1983; Nazareth, 1989). Learning from example (LFE) paradigms, on the other hand, do not require the experts to specify how they make their decisions. When applied to ES, LFE models function as *implicit knowledge acquisition* (IKA) tools, where the required knowledge is implicitly embedded in the examples. For this reason it is relatively easy to show that ML-based expert systems are usually less expensive than their EKA-based counterparts. But what about the relative performance? Most importantly, do the cheaper construction costs of ML-based expert systems imply poor accuracy?

There is currently very little scientific evidence of experiments which compare ML models vs. EKA methods in terms of classification accuracy. Although ML models have been

successfully used in an increasing number of real-world applications, most of the reports so far compare the performance of various LFE models with statistical models (for example: Tam, 1992; Thrun et al., 1991), but not with EKA methods. An exception to this statement can be found in the pioneering work which was done by Michalski and Chilausky (1980). To their surprise, M&C, have found that the AQ11 ML model did better in terms of classification accuracy than the explicit rules they derived from domain experts. However, as far as we know, their experiment was not repeated in different contexts, such as by using other machine learning models, different problem domains, and other subjects. Furthermore, the AQ11 in M&C's experiment included prior domain knowledge, which was derived from experts. Consequently, M&C's comparison was actually between a ML model which was already using EKA techniques versus a pure EKA method. We are interested mainly here in comparing LFE methods which do not include any prior domain knowledge with EKA methods in terms of classification accuracy. The main contribution of this experiment is the provision of more evidence that ML can well compete with EKA in terms of classification accuracy. In this respect, the two ML models, neural networks and the OLM, which were used in this experiment, outperformed the explicit rules which were provided by the subjects. The experiment and its results can, thus, assist in resolving the debate whether ML can offer anything to expert systems.

2. The experiment

The subjects of the experiment were undergraduate Business Administration students. They were presented with a problem they all have been familiar with—lecturer evaluation. In Part I of the experiment, the students were asked to grade hypothetical lecturers. In Part II they were requested to explicitly express how they have made these grades. The lecturers' attributes were: (A) Teaching of important concepts and tools, (B): Teaching methodology, (C): Ability to raise interest of students, and (D): Attitude towards the students. The four selected attributes summarized a more detailed questionnaire which is routinely distributed among all the students at the conclusion of every course they attend. The subjects were not experienced decision-makers. However, as mentioned above, they all were experienced in rating lecturers. To simplify the problem domain, each attribute, as well as the final grade, had only five possible ordinal (i.e., ordered) symbolic values.

In Part I, the students were requested to assign final grades to one hundred hypothetical lecturers, whose attribute values were selected at random beforehand. In Part II of the experiment, it was recommended that the students use IF ⟨attribute values⟩ THEN ⟨final grade⟩ rules. One site assistance was available during the whole experiment. Neither Part I nor Part II were limited in time. Participation was voluntary, and anonymity was strictly kept. The fact that the lecturers were hypothetical may have been working against the subjects. However, they were not judged against any normative decision-making model. The subjects were only requested to match the rules in part II with their own actual classifications of Part I. The purpose of the experiment was explained to the students both orally and in writing. The task was presented as an intellectual challenge. The students were told that usually it is not an easy task to match the rules of Part II with the judgements they have made in Part I (Larichev et al., 1988). They were encouraged, however, to try and achieve this

goal as much as they could. From the sixty students who heard the initial explanations, 39 participated in Part I. Only 14 of them were filling out Part II of the questionnaire to their satisfaction. The findings of this research, to be shortly presented, refer to those fourteen students. Most of the subjects finished Part I in less than twenty minutes. The last one finished in half an hour. Part II required more than an hour for most subjects. The last one finished Part II in about ninety minutes.

The questionnaires of Part I and Part II were processed as follows: The attribute-score pairs of Part I were randomly partitioned to training sets (70% of the examples) and holdout samples (30%). Fourteen pairs of training sets and holdout samples were, thus, available. Back propagation neural networks (Rumelhart, 1986) were selected to represent non-symbolic LFE models, since they are well known to provide relatively good classification accuracy (Tam, 1992: Thrun et al., 1991). The training sets were used to train fourteen feed forward neural networks. In order to accelerate convergence, nonlinearity was introduced at the input level through the introduction of second order cross products and periodic functions of inputs (Details of this method can be found in the description of N-Net, 1990). Once the neural networks were trained, they were used for predicting the respective holdout samples.

The same training sets and holdout samples were used by the OLM (Ben-David et al., 1989; Ben-David, 1992), which is an exemplar-based LFE model (Kibler & Aha, 1987). The OLM was used since all the lecturers' attributes, as well as the final grades, were ordinal symbols. Unlike neural networks and TDIDT algorithms, (Quinlan, 1983; 1987), the OLM uses the order within the domains. Neural networks enforce numeric scales on ordinal symbols, and TDIDT models ignore them altogether. The OLM's predictions are consistent with each other, in the sense that monotonicity among subsequent classifications (with respect to each attribute) is always kept. This property prevents situations in which a lecturer all of whose attributes are better than another's will receive a lower grade during classification. Fourteen regression equations were also derived using the training sets, and predictions were made on the holdout samples to serve as a benchmark.

3. Results

The attribute data were composed of integers (ranging from 0 to 4) with mean close to 2 and standard deviation of 1.45. The mean score of the lecturers performance was 1.75 with standard deviation of 0.99. The consistency within the subjects' judgements was also checked in the sense of monotonicity. Each attribute-score pair was compared with all the other pairs which were provided by the same subject. An attribute-score pair was considered inconsistent with another if all its attribute values were higher or equal (lower or equal) than those of the other, while its score was strictly lower (higher). More details can be found in Ben-David et al. (1989; 1992). Each subject classified 95 lecturers on the average. The average number of inconsistent attribute-score pairs was 84.57 (per subject) with standard deviation of 21.04. This number is quite reasonable, considering that $4465((95^2 - 95)/2)$ pairwise comparisons are needed for checking the consistency of 95 attribute-score pairs (see also Larichev et al. (1988)). The average inconsistency rate detected in Part I was 1.92 percent with standard deviation of 0.34. The average number of rules provided by each candidate was 12.79 with standard deviation of 5.25. The average inconsistency rate

Table 1. Mean square error of various IKA and EKA methods.

Subject	Neural Networks	OLM	Regression	Explicit Rules
1	0.159	0.300	0.133	0.467
2	0.384	1.200	0.333	2.233
3	0.403	0.733	0.367	0.677
4	0.490	0.500	0.567	1.600
5	0.505	0.500	0.567	2.068
6	0.466	0.433	0.167	1.000
7	0.165	0.533	0.333	1.800
8	0.345	0.433	0.300	0.677
9	0.256	0.933	0.333	1.933
10	0.405	0.800	0.600	2.733
11	0.569	0.100	3.400	0.967
12	0.115	0.367	0.100	0.800
13	0.221	0.700	0.267	0.600
14	0.305	0.733	0.533	1.500
Average	0.319	0.551	0.533	1.269
Variance	0.025	0.090	0.618	0.560

of the rules, checked as explained above, was 4.38 percent with standard deviation of 4.81. Clearly, in this respect, the actual classifications of Part I were more consistent with each other than the rules which were expressed in Part II.

Table 1 shows the mean square error (MSE) of the predictions over the holdout samples for each subject.

Figure 1 shows the cumulative distributions of the holdout samples MSEs.

A paired T test of the MSE distributions is given in Table 2 with their corroesponding 2-tailed probabilities.

4. Discussion and conclusions

The results of the empirical experiment show that the LFE models used here, both symbolic and numeric, had better classification accuracy than the explicit rules. Although significant statistical advantage of LFE models over explicit knowledge acquisition was demonstrated here, major simplifying assumptions were made:

(1) The problem domain was simple, with only four attributes, and with no clear demonstration that expertise was actually needed. This fact could have been biased the results in favor of the ML models. (2) Both the actual classifications, we well as the set of rules which represented EKA-based expert systems referred to hypothetical lecturers. This could have been the reason why the subjects did not spend more time on writing down more accurate rules. (3) The subjects were undergraduate students. Although they were all experienced with actual lecturer ratings, they are clearly cannot be considered as experienced decision-makers. (4) The first version of the rules was considered final. This is unlike typical expert systems which are built interactively over a long period of time.

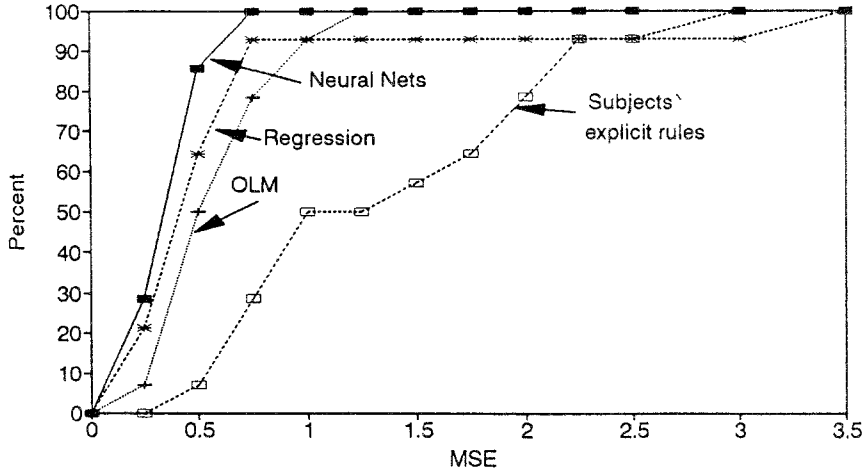

Figure 1. Cumulative distribution of MSE of various IKA and EKA methods.

Table 2. Pairwise T tests of MSE distributions.

	Neural Networks	OLM	Regression	Explicit Rules
Neural networks	—	$2.85; p < .014$	$1.13; p < .279$	$5.50; p < .001$
OLM		—	$0.07\ p < .944$	$4.78; p < .0001$
Regression			—	$2.26; p < .020$
Explicit rules				—

The number of subjects was also quite modest. So was the number of examples each subject provided. For these reasons, caution must be taken while interpreting the results of this experiment. It will be incorrect, for example, to conclude from these results that ML models are always superior to EKA methods. More empirical research is clearly needed.

However, were the results of this experiment expected? It is well known that humans poorly solve multiattribute decision-making problems. They are typically very biased (see Tversky & Kahneman, 1985; Einhorn, 1971; Ganzach, in press), and resort to various simplifying problem-solving techniques. The fact that linear (or almost linear) regression models human decision-making quite well is also well known. It was reported by decision-making researchers during the Seventies (Dawes, 1974; Wainer, 1976). Furthermore, machine learning models have been often shown to be more accurate then well established statistical techniques (Tam, 1992; Thrun et al., 1991), in particular, when compared with various regression models. One could, therefore, expect ML to excel in the modeling of human decision-making. However, we think that more scientific evidence is net needed. In particular, in relation with the comparison between the prediction accuracy of ML vs. EKA methods.

References

Ansari, A., & Modarress, B. (1990). Commercial use of expert systems in the US. *Journal of Systems Management*, December 1990, 11–16.

Ben-David, A., Sterling, L., & Pa, Y.H. (1989). Learning and classification of monotonic ordinal concepts, *Computational Intelligence*, 5 (1), 45–49.

Ben-David, A. (1992). Automatic generation of symbolic multiattribute ordinal knowledge-based DSSs: Methodology and applications. *Decision Sciences*, 23(6), 1357–1372.

Buchanan, B.G. (1989). Can machine learning offer anything to expert systems. *Machine Learning*, 4(3/4), 251–254.

Dawes, R.M., & Corrigan, B. (1974). Linear models in decision-making. *Psychological Bulletin*, 81, 95–106.

Einhorn, H.J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Psychological Bulletin*, 73, 221–230.

Ganzach, Y. (in press). Goals as determinants of nonlinear noncompensatory judgment strategies: Leniency versus strictness. *Organizational Behavior and Human Decision Processes.*

Hayes-Roth, F., Waterman, D.A., & Lenat, D.B. (1983). *Building Expert Systems.* Addison-Wesley.

Kibler, D., & Aha, D.W. (1987). Learning representative exemplars of concepts: An initial case study. In P. Langley (Ed.), Proceedings of the 4th International Workshop on Machine Learning. Morgan Kaufmann Publishers, Los Altos, CA, 24–30.

Larichev, O.I., & Moshkovich, H.M. (1988). Limits to decision-making ability in direct multiattribute alternative evaluation. *Organizational Behavior and Human Decision Processes*, 42, 217–233.

Michalski, R.S., & Chilausky, R.L. (1989). Knowledge acquisition by encoding expert rules versus computer induction from examples: A case study involving soybean pathology. *International Journal of Man-Machine Studies*, 63–87.

Nazareth, D.L. (1989). Issues in the verification of knowledge in rule-based systems. *Int. J. of Man-Machine Studies*, 30, 255-271.

N-Net User Manual Ver. 2. (1990). AI Ware Inc., University Circle, Cleveland, Ohio 44106 USA.

Quinlan, J.R. (1983). Learning efficient classification procedures and their applications to chess and games. In R.S. Michalski, J. Carbonell, & T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Co., San Mateo, CA: Morgan Kaufmann.

Quinlan, J.R. (1987). Simplifying trees, *Int. J. of Man-Machine Studies*, 221–234.

Rumelhart, D.E. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. Mclelland (Eds.), *Parallel Distributed Processing, 1*, The MIT Press.

Tam, K.Y., & Kiang, M.Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. Management Science, 38 (7), 926–947.

Thrun, S.B. et al. (1991). The Monk's Problems: A performance comparison of different learning algorithms. Computer Science Dept. Carnegie Mellon University, CMU-CS-91-197.

Tversky, A., & Kahneman, D. (1985). Judgement under uncertainty: Heuristics and biases. In *Judgement under Uncertainty: Heuristic and biases*, D. Kahneman, P. Slovic, & A. Tversky (Eds.), Cambridge University Press.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no never mind. *Psychological Bulletin*, 85, 213–217.

Received June 1, 1993
Accepted June 1, 1993
Final Manuscript June 8, 1994