Empirical Analysis Of The General Utility Problem In Machine Learning

Lawrence B. Holder
holder@cse.uta.edu

Department of Computer Science Engineering University of Texas at Arlington Box 19015, Arlington, TX 76019-0015

The utility problem in speedup learning and the overfit problem in inductive learning show a common behaviour of the machine learning methods i.e.peak attainment of performance followed by gradual decrease in performance due to increasing amount of learned knowledge.This is known as the *general utility problem* in machine learning.

Initially [4] used the name *utility problem* as a description of the problem in speedup learning but later it was generalized to other machine learning paradigms.

Overfit occurs when the learning method starts to learn the outliers/noise or learns the data too well.Hence,errant patterns may be due to noise in the training data or inadequate stopping criteria of the method. From Figure 1 , we notice that a model of this kind can be used to avoid the performance degradation by controlling the amount of learned knowledge to coincide with the peak of the performance response.

[2] described MBAC(Model-Based Adaptive Control) system that uses an empirical model of the performance response to control learning.Also experiments with MBAC show that the parabolic model is capable of choosing an appropriate learning method and controlling that method[2] but MBAC suffers from inaccuracies due to differences between empirical and true model of performance response. So, MBAC or any approach to controlling and estimating the performance of a learning method would benefit from formal model of the performance response which depends on properties of current learning task .

It is observed that incase of splitting method , neural network learning method and set covering method each on of them follow the general utility problem trend due to their overfit behaviour.

In Splitting methods,when plotting no. of split vs the classification accuracy of the knowledge after each split.We, observe that for ID3 inductive learner on different domains, the performance response of these follows the figure1.(A) trend i.e. first , they reach a peak and progress with a gradual decrease.The knowledge produced by splitting method can be represented as a decision tree and when using different traversal methods, we see that the effect of overfit becomes more significant as decision tree starts to go deeper.Therefore, the breadth first traversal has overfit at later stages of splits as compared to depth first and best first traversal method. On the flag domain, such an trend as in figure1(A) is spotted.Using chi-square pre-pruning and reduced-error post-pruning help reduce the overfit problem but the tree still has a lesser accuracy than peak accuracy of the performance response.Also,When using PLSI method, such an result was obtained as given in figure 1.(C).So, we can say that these pruning techniques do not completely alleviate the overfit problem.

In the set covering method, as we generally learn DNF(disjunct normal form) expressions for the hypotheses.Hence, the dimension used to measure the amount of learned knowledge is the number of disjuncts in the hypothesis induced.[3] compared the accuracy of the complete DNF hypothesis produced by AQ to truncated versions of the same hypothesis. The hypothesis consists of the single disjunct covering most examples (best disjunct). The second truncated hypothesis has disjuncts covering more than one unique example i.e. (unique > 1). The truncated hypotheses use a simple matching procedure to classify the multiply-covered and uncovered examples. Seeing figure 2.(B), we conclude that for all the medical conditions, a near trend of figure 1.(A) is followed.

In neural network methods.As the number of cycles increases, training instances are more accurately classified.But over the increasing cycles ,overfit starts to occur as the network learns the training instances too well and degrades the accuracy.So, here also, trend of figure 1.(A) is followed.so,we see that we can reduce the worsening of the performance of learner by limiting the amount of learned knowledge to the the point corresponding to peak performance.

The performance response model offers a general method for avoiding the general utility problem in many machine learning methods and has a shape that is result of bias-variance trade-off.The number of leaves L in the decision tree expresses the bias and variance in the analysis. Assuming binary splits at each node, L - 1 is the no. of splits.So,as L increases, we can say, similar will be behaviour of bias variance when no. of splits increases.

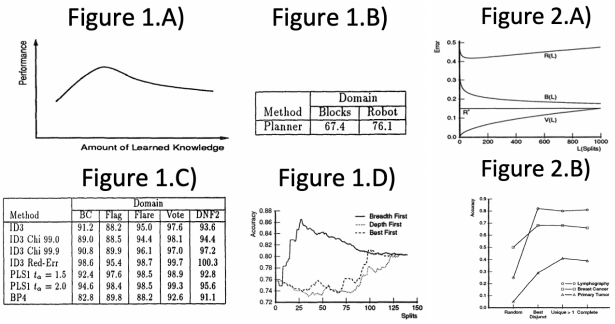

Figure 1: *(A)general performance response of a learning method that has general utility problem.The units along the horizontal axis represent a simple transformation in the learner's hypothesis increase along this axis is a refinement in the existing knowledge. The vertical axis measures the performance of the learned knowledge after each transformation.(B)Entries for the Planner speedup learner on two different domains(percentage)(C)Percentage final performance of peak for inductive learners.(D)Performance of on the DNF2 domain of ID3 for three different decision tree expansion.*

Figure 2: *(A)Performance response as for a decision tree induction method. (B)Performance response of AQ for various medical problem and also different disjunct methods.*

The expression for the classification error $R(L)$ in terms of the bias $B(L)$ and variance $V(L)$ is given where R* is the Bayes optimal classification error is given as:

$$R(L) = B(L) + V(L) + R^* \quad (1)$$

$$V(L) \leq \sqrt{L/N}, B(L) \leq C/L^{2/M}, V(L \approx N) \leq R^* \quad (2)$$

Here, C is a constant, M is the dimension of the instance space (i.e., number of features used to describe the training instances) and N is the number of training instances.Also, Equation 1 is an expression of the classifiction error response curve.

Subtracting this error curve from Equation 1 would yield the accuracy response curve.This performance response being similar to that of Figure 1.(A) supports the existence of a single peak and the inevitability of overfit in splitting algorithms without appropriate stopping criteria or post-pruning techniques.

So, we see that the speed up learning,splitting method , neural network learning method and set covering method, each on of them follow the trend of figure 2.(A) and therefore, as discussed earlier, all of these depict a common trend as in figure 1.(A).A Model of this trend can be used to control the amount of learned knowledge to achieve peak performance and also predict the achievable performance of the learning method as a means of selecting an appropriate method for a learning task.Hence from all of the above, we conclude that the forces of bias and variance and the constraints on the order of knowledge transformations serve to bring together several methods and that the continued refinement of models of the general utility problem will provide a general framework for controlling and comparing different learning paradigms .

[1]Lawrence B. Holder;1992;Empirical Analysis Of the General Utility Problem In Machine Learning

[2]Holder, L. B. 1991a. Maintaining the Utility of Learned Knowledge Using Model-Based Adaptive Control.

[3]Michalski, R. S. 1989. How to learn imprecise con- cepts: A method based on two-tiered representation and the AQl5 program. In Kodratoff, Y. and Michalski, R. S., editors 1989, Machine Learning: An Artificial Intelligence Approach, Vol III. Morgan Kaufmann Publishers.

[4]Minton, S. 1988. Learning Search Control Knowledge: An Explanation-Based Approach. Kluwer Academic Publishers.