On the Importance of Change of Representation in Induction

Eduardo Pérez, **Ricardo Vilalta** and **Larry A. Rendell**
perez@cs.uiuc.edu vilalta@cs.uiuc.edu rendell@cs.uiuc.edu
Beckman Institute and Department of Computer Science, Univ. of Illinois
405 N. Mathews Avenue, Urbana, IL 61801, USA

Abstract

The adequacy of current machine-learning techniques relies on an implicit built-in assumption in which simplicity and similarity play significant roles. This paper discusses on the insufficiency of these assumptions for domains where complex attribute interactions characterize the target concept (i.e. on difficult domains), and revives a model of induction where change of representation, through the construction of increasingly more abstract knowledge structures, contributes to increased learning performance. We describe two learning systems proved effective on difficult domains, their success attributed to the use of knowledge structures that bridge the gap between primitive and complex representations. Our conclusions emphasize the importance of change of representation as a fundamental step during inductive learning.

Background: Success and failure in inductive learning

Past successes and failures of learning techniques have been conditioned by an assumption implicit in a model of induction shared by many researchers in machine learning, neural networks, pattern recognition, genetic algorithms and statistics. Current learning techniques often rely strongly on simplicity and similarity, which can be both useful and harmful. Early emphasis on simplicity and similarity (Rendell, 1986) allowed researchers to progress and produce systems that, when used in the appropriate industrial applications, have generated savings of millions of dollars (Langley & Simon, 1995). However, applications such as protein structure prediction still are beyond the capabilities of current learning methods (Ioerger, Rendell, & Subramaniam, 1995).

Attempts to improve learning methods have produced only marginal progress on problems where performance was already good. Difficult domains, involving hard concepts (Rendell & Seshu, 1990), have their structure hidden by poor representations. In practice, we expect to confront primarily problems that are structured, not random (Abu-Mostafa, 1988). Considering structured and random concepts under uniform distribution may produce misleading results (Schaffer, 1994; Rao, Gordon, & Spears, 1995). Still, structured concepts are difficult for current learning methods when only primitive attributes are used to describe the data. Primitive attributes must be used when experts lack domain knowledge to define better attributes, more informative and predictive ones. An abundance of primitive attributes exacerbates attribute interaction, and complicates learning. When attribute interaction is high, the target function is rough, with many peaks of small size, complicated shape, and perplexing arrangement throughout instance space (Rendell & Seshu, 1994). In such situation, using instance proximity and hypothesis simplicity is questionable, as expressing hypotheses simply in terms of only primitive attributes is inappropriate. Each individual attribute carries little information, so it is difficult to choose a good one. Also, hypothesis descriptions are so large that constructing them by incrementally adding individual primitive attributes requires many unreliable steps of composition.

Functional requirements of learning: toward a model of induction

Progress in inductive learning depends on the ability to learn complex structured concepts effectively. Defining learning as the discovery of a function (Natarajan, 1989; Rendell & Seshu, 1990) emphasizes the idea that nothing is known about the target function (except the training data). That is, the form or model of the function also needs to be discovered. Like function approximation, inductive learning means more than parameter estimation; it means constructing the model itself. This difficult goal has taken several forms in different research fields, including change of representation, variable bias, adaptable network structure, and non-parametric statistics (Schlimmer, 1987; Utgoff, 1986; Rendell, Seshu, & Tcheng, 1987; Barron & Barron, 1988; Devroye & Gyorfi, 1985; Geman, Bienenstock, & Doursat, 1992).

Induction involves the discovery of regularities in data and knowledge structures. The complexity of the existing regularities determines the complexity of the inductive process. Whenever structure is intricate, regularities are involved, and it is useful to view induction as a multi-layered process (Rendell, 1985). Successive layers form knowledge structures more abstractly. The discovery of progressively more complex structures is accompanied by an increased compression of class-membership information. Ultimately, the goal is to obtain an inductive representation that elucidates all the mediating structures from primitive attributes to the target function (Rendell, 1987).

One primitive model of induction is based on three levels of knowledge structures as defined by Rendell (1985). At the lowest level, compressing class-membership information from neighboring points in instance space leads to formation of patterns (or clusters). Next, similar patterns coalesce into pattern classes (e.g., relations—see below), merging possibly distant, but homogeneous regions of instance space. The highest level uses transformation operators (induced or suggested by domain knowledge) to create pattern groups, which can predict unobserved instance space regions by transformation of similar structures discovered at lower levels. Each level involves a different kind of similarity, but current inductive systems can only observe similarity at the lowest level, which precludes learning in difficult domains.

Case studies: Relational and hierarchical representations

Guided by the need for change of representation in learning, we have recently developed two systems that improve upon previous ones, especially when the focus is on complex structured concepts. First, an approach called MRP emphasizes finding relations as a way to induce better representations for learning, and uses the algebraic notions of relation and relational operators to construct hypothesis based on relations extracted from the data. The second approach, HFC (Vilalta, Blix, & Rendell, 1996), focuses on building a hierarchy of features (e.g., unrestricted Boolean formulas) which identifies progressively more complex knowledge structures as higher levels of the hierarchy are formed. Both approaches induce knowledge structures and use them to learn the final target more accurately.

For MRP, whose central operator is **M**ultidimensional **R**elational **P**rojection, the training examples with the same class are a relation, i.e., a set of tuples. MRP projects data relations onto subspaces of decreasing size, while searching for a projection that adequately balances information loss and data compression. This search is repeated until all data are explained by MRP's current hypothesis, composed of relations found in the data. Pérez & Rendell (1995) showed that MRP learns realistic complex concepts more accurately than six other learning systems, including a sophisticated system for rule induction and two advanced methods for feature construction. Forty Boolean concepts were used to evaluate MRP in difficult learning contexts (where entropy and concept dispersion were high). When expressed in DNF using only primitive attributes, these concepts involved hundreds of terms and thousands of literals. However, the same concepts could be expressed concisely as combinations of complex relations, such as parity, M-of-N, majority, and others frequently used to evaluate learning techniques. As an illustration, using 205 training examples (i.e., 5% of the instance space), MRP's accuracy (averaged over the 40 concepts) was 8.1 percentage points higher than that of its best competitor selected on a concept-by-concept basis. In addition to Boolean concepts inspired by realistic applications (such as synthesis of combinational circuits), we also considered real-world applications. With real-world datasets, MRP's accuracy advantage over other systems was more noticeable when data representations were primitive, not carefully manufactured by experts. MRP learned more accurately than the other systems two important concepts in secondary protein structure prediction (Qian & Sejnowski, 1988). MRP's advantage on this domain can be traced back to role of complex relations in human-made theories of protein folding (Chou and Fasman, 1974). Finding unrestricted multivariate relations in the data becomes fundamental to learning despite attribute interaction and concept variation (Pérez & Rendell, 1996).

HFC (for **H**ierarchical **F**eature **C**onstruction) builds a hierarchy of increasingly complex Boolean formulae in a bottom-up fashion, by progressively incorporating new levels of abstraction on top of previous levels. The most basic layer corresponds to all primitive boolean features, intermediate layers to new formulae/features constructed by the application of boolean operators, and the the top feature in the hierarchy denotes the output hypothesis.

A new layer L_i is formed from the union of two feature sets, $L_i = {L_i}^r \cup {L_i}^c$. The first set, ${L_i}^r$, proceeds by individually *specializing* each of the features in the layer below, L_{i-1}. This operation generates new features that improve upon the identification of the multiple single-class regions in the instance space. The second set, ${L_i}^c$, looks *globally* to all previous layers, $L_1, L_2, ..., L_{i-1}$, by searching for new feature *combinations*. The goal here is to coalesce dispersed single-class regions to provide a transformation of the instance space (i.e. for effectively producing a change of representation). This two-step process is repeated to generate new layers until either a consistent feature is found, or the best feature at L_i cannot improve over the best feature at L_{i-1}. Each new feature layer comprises more specialized features (set ${L_i}^r$), and new feature combinations (set ${L_i}^c$). The idea is to progressively smooth the original instance space to ease the distinction between positive and negative examples. The top of the hierarchy is occupied by the the final concept estimation.

HFC is the result of an investigation aimed at extracting design elements we regard effective in building feature-construction algorithms. By studying/testing several well-known learning methods we identified key factors contributing to the effectiveness of discerning the concept's nature. Such design elements are usually mingled with assumptions that tend to degrade classification performance. Preliminary tests on HFC have shown significant improvement in terms of predictive accuracy when compared over other learning techniques. Further analysis/experimentation of HFC's main components will provide important insights for effective change of representation during inductive learning.

The role of knowledge

One way of viewing the problem of finding good knowledge structures uses a standard AI paradigm: search. The constructive induction system searches a space of structures. To evaluate a node, we employ some criterion, which can be defined solely in terms of some semantic assessment and accuracy measurement. But if any domain knowledge is available, it may be converted into a further means of limiting search. This is the subject of Donoho's (1995, 1996) work, which studies the various effects of knowledge on the instance space and structure space, categorizes kinds of knowledge, and tests principled methods for using (and improving) it. Ioerger (199?) uses a related approach for search using knowledge in more general domains.

Discussion

Systems based on flexible representation schemes learn hard concepts more accurately than other sophisticated current methods. We present this as evidence that it is possible to extend the scope of practical applications where inductive learning can succeed. The partial success of our case studies is attributed to focusing the learning process on the construction of intermediate knowledge structures (relations and hierarchical features) as a new representation. The similarity- and simplicity-based model of induction implicit in much of past research needs to be changed before significant progress can be attained. There are situations in which hypothesis simplicity and instance similarity are not appropriate biases. Whenever possible, we must develop inductive learners that create their own representations as part of their learning. Change of representation is not just a technique for learning. Constructing new (inductive) representations during learning is a fundamental goal of learning itself.

References

Abu-Mostafa, Y. S. (1988). Complexity of random problems. In Abu-Mostafa, Y. S. (Ed.), *Complexity in Information Theory*, chap. VI, pp. 115–131. Springer-Verlag, NY.

Barron, A. R., & Barron, R. L. (1988). Statistical learning networks: A unifying view. In *Proceedings of the Symposium on the Interface: Statistics and Computing Science* Reston, VA.

Devroye, L., & Gyorfi, L. (1985). *Nonparametric Density Estimation: The L1 View.* Wiley, New York, NY.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Ioerger, T. R., Rendell, L. A., & Subramaniam, S. (1995). Searching for representations to improve protein sequence fold-class prediction. *Machine Learning*, *21*, 151–175.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, *38*(11), 54–64.

Natarajan, B. K. (1989). On learning sets and functions. *Machine Learning*, *4*, 67–97.

Pérez, E., & Rendell, L. A. (1995). Using multidimensional projection to find relations. In *Proc. of the Twelfth Int. Conf. on Machine Learning*, pp. 447–455. Morgan Kaufmann Pub., Inc.

Pérez, E., & Rendell, L. A. (1996). Learning despite concept variation by finding structure in attribute-based data. In *Proceedings of the Thirteeth International Conference on Machine Learning.* Morgan Kaufmann Pub., Inc.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural networks. *Journal of Molecular Biology*, *202*, 865–884.

Rao, R. B., Gordon, D., & Spears, W. (1995). For every generalization action, is there really an equal and opposite reaction? analysis of the conservation law for generalization performance. In *Proc. of the Twelfth Int. Conf. on Machine Learning*, pp. 471–479. Morgan Kaufmann Pub., Inc.

Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*, *1*(2), 177–226.

Rendell, L., Seshu, R., & Tcheng, D. (1987). Layered concept-learning and dynamically-variable bias management. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pp. 308–314 Milan, Italy.

Rendell, L. A. (1985). Substantial constructive induction using layered information compression: Tractable feature formation in search. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 650–658.

Rendell, L. A. (1987). Conceptual knowledge acquisition in search. In *Computational Models of Learning*, pp. 87–159. Spring-Verlag, New York.

Rendell, L. A., & Seshu, R. (1990). Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence*, *6*, 247–270.

Rendell, L. A., & Seshu, R. (1994). Learning hard concepts through constructive induction: Framework and rationale. In Hanson, S. J., Drastal, G. A., & Rivest, R. L. (Eds.), *Computational Learning Theory and Natural Learning Systems*, Vol. I, chap. 5, pp. 83–141. MIT Press, Cambrigde, MA.

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 259–265. Morgan Kaufmann Pub., Inc.

Schlimmer, J. C. (1987). Learning and representation change. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pp. 511–515.

Utgoff, P. E. (1986). Shift of bias for inductive concept learning. In Michalski, R., Carbonell, J., & Mitchell, T. (Eds.), *Machine Learning: An Artificial Intelligence Approach, Vol. II*, chap. 5, pp. 107–148. Morgan Kaufmann Publishers, Inc., Los Altos, CA.

Vilalta, R., Blix, G., & Rendell, L. A. (1996). Hierarchical feature construction in inductive learning. Submitted to *the Journal of Artificial Intelligence Research.*