Decision Forest Enhancement Evaluation

Thomas E. Sandman
School of Business, CSUS, sandmant@csus.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis1995

Recommended Citation
http://aisel.aisnet.org/amcis1995/64

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 1995 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Specific Aims

This research involves a cross fertilization of Computer-Assisted Decision Making in Business and Artificial Intelligence (AI) in Computer Science. This project represents a third phase of an ongoing research agenda exploring the concept of an 'Induction Support System' which would be used in the process of using induction to develop knowledge-based systems. This research project focuses on comprehensive empirical research that is necessary to assess the impact of two proposed enhancements to a normative approach for the use of a significant machine learning technology [4].

There are two evaluative components of an induced knowledge base. The first is the accuracy of the knowledge. The second is intelligibility of the knowledge base. This is the degree to which an induced knowledge base can be comprehended by someone knowledgeable in the related domain. This is extremely important. Induction algorithms are data driven, i.e. the resultant knowledge base has been created based on the attribute values, not on the semantic nature of the attributes themselves. Prior to implementing and using an induced knowledge base, one must attempt to attach construct validity (i.e., semantic meaning) to the knowledge base [3].

The goal of this research is to evaluate two specific normative enhancements to the decision forest approach to using a specific class of induction algorithms. These two enhancements will help build knowledge bases which: 1) are accurate, and 2) provide greater opportunity for construct validation. This research evaluates a novel approach to deriving knowledge through the use of induction. This research represents an ongoing exploration of heuristics for using top-down induction of decision tree algorithms begins to address a gap in the literature [4, 5, 6, 7, 8, 9].

Background and Significance

Over the past few years, designing and implementing Knowledge-based Decision Support Systems (KBDSS) that incorporate AI technology have represented an important new research arena. Much has been learned in the past decade related to preferred KBDSS development methodologies, predictable pitfalls, desirable technological approaches, etc. The acquisition of knowledge for inclusion into a KBDSS, however, remains an under-addressed research issue.

Researchers recognize the "knowledge acquisition bottleneck" constrains wide-scale development of practical systems. AI research focusing on machine learning algorithms
provides an interesting platform from which novel tools can be derived to address the knowledge acquisition bottleneck problem. An indirect outcome of machine learning research has been the observation that such technology provides a promising approach for lightening the knowledge acquisition burden. Unfortunately, there is little empirical evidence to substantiate this observation. This issue is further clouded by the large number of machine learning algorithms that have been proposed and/or implemented. For example, neural networks are popular in the current literature, yet they yield 'black box' knowledge bases and fail to provide opportunity for construct validity [3].

This research describes two enhancements to a 'decision forest' approach for using a specific class of inductive machine learning algorithms to assist developers in alleviating the knowledge acquisition bottleneck problem. Once validated, these enhancements could be incorporated into a tool designed to facilitate a comprehensive KBDSS knowledge acquisition research agenda.

The most widely used algorithm in available induction systems is Quinlan's Iterative Dichotomiser 3 (ID3). Variations of ID3 are often addressed in the artificial intelligence literature, and the ID3 approach is exemplary of other existing top-down induction of decision trees (TDIDT) techniques. TDIDT systems use a set of test cases, e.g., \{t_1, t_2, \ldots, t_i, \ldots, t_n\}, selected because they are representative of some population (W). Each example, \(t_i\), consists of a 'class designation (c)' followed by a series of 'attributes (a)' predicted to describe that class designation in W, e.g., \(t_i=cp, a_1, a_2, a_3, a_n\). From a set of test cases, induction systems generate 'decision trees' that classify all of those test cases.

**Figure 1. Typical Use of Induction Systems.**
Typical induction systems have been designed to build a single decision tree for a specific run. As shown in Figure 1, domain experts and/or knowledge engineers gather a set of training examples representative of some population (W). This training set is then input to the learning algorithm. An ID3-like algorithm will produce a decision tree which is incorporated into a knowledge-base as a set of rules. Once the performance of the system has been validated by the experts, users can query the system for advice concerning the domain related classification problem. The user should also learn more of the domain through the use of the KBDSS.

TDIDT induction systems are the best for use in a comprehensive KBDSS knowledge acquisition research agenda [5, 6, 7, 8, 9]. Seven reasons support this: they enhance the articulation of expertise, they provide unbiased perspectives, they integrate multiple expertise, availability, efficiency of the TDIDT algorithms, ease of prototyping KBDSS, and the maturity of related research.

Project Design and Methodology

As stated above, the goal of this research is to evaluate two specific normative enhancements to the decision forest approach to using a TDIDT induction algorithms. The first enhancement modifies the current voting in the decision forest approach by weighting each rule by the number of training cases that that rule is consistent with. This would give the rules which appear to be more generalizable a greater weight in the knowledge base. The second enhancement is independent of the first. It examines a 'pruning' approach to be used for the decision forest. This pruning enhancement would eliminate rules from the forest which have been produced to identify a single training case from the training set. This should reduce the potential for generalizing rules from potential outlier training examples.

The general methodology followed in the machine learning literature is to define the training and test data, apply the technique, and evaluate the results. Data sets representing different domains are divided into training and validation sets of examples. Training sets of examples are used to create knowledge bases, and these knowledge bases are evaluated with validation sets of holdout examples. Comparative evaluation for accuracy and intelligibility would be conducted.

The specific methodology is as follows:

1. Define Data Sets.

   The Braun and Chandler [1] data set and the Mingers [2] data sets have been secured, which provides a valuable link to the related body of literature.

2. Define Training Sets.

   For each data set, a 10-fold cross validation approach will be used with random assignment to the 10 training sets extracted from each data set.
3. Create the Knowledge Bases.

Each of these 50 training sets will be used to create three different knowledge bases. These are: 1) The typical TDIDT decision tree, 2) The decision forest [5,6], and a pruned decision forest using the enhancements described in step 4.

4. Apply the enhancements.

Appropriate weights need to be calculated and incorporated into the decision forest knowledge bases. This would yield the weighted decision forest (WDF) knowledge bases. All branches (or rules) which identify only one training example would be removed from the knowledge base. This would yield the pruned decision forest (PDF) knowledge bases.

5. Evaluate the Knowledge Bases.

Ten-fold cross validation will be used. The classification accuracy will be measured and metrics for intelligibility will be used (e.g., # of rules in the databases, and # of cues per rule).


The results for each data set will be averaged as is done in the literature [1,2]. Furthermore, the performance of each of the enhancements will be measured across domains (using least significant difference measures of t-tests at varying a levels).

7. Write Papers.

Submit the results of the study to related journals.

Limitations

One of the major benefits of using this approach is to generate a large variety of rules in a short time. This variety should greatly enhance the prospects of arriving at a knowledge base that is accurate and that has construct validity. One major strength of this project is the reliance on data sets that have been previously used in major research efforts. While this enhances this research, it prevents the true assessment of construct validity by this researcher. Therefore, the proxies for intelligibility are used. This is based on the assumption that it is easier to assess the construct validity for a 'more intelligible' rule than a 'less intelligible' rule.

Acknowledgments

This research is being sponsored by the CSU system through a State Funded Research, Scholarly and Creative Activity Summer Fellowship Award.
References


