Submission Cover Sheet-IEEE Southeastcon '98

Submission Type (Due Date): Concise Paper (12/15/97)
Title of Paper: A Comparison of Classification Algorithms on Highly Nonlinear Signal Processing Problems

Authors and Affiliations (List authors in the order to appear on the manuscript. List affiliations (companies, universities, etc.) and location exactly as they should appear. If there are additional authors, attach a second sheet.

1. **Name:** Audrey Le
Affiliation: Inst. for Signal and Information Proc.
and Dept. of Elect. and Comp. Engr.
Location: Mississippi State, MS 39762

2. **Name**: Julie Ngan
Affiliation: Inst. for Signal and Information Proc.
and Dept. of Elect. and Comp. Engr.
Location: Mississippi State, MS 39762

3. **Name:** Janna Shaffer
Affiliation: Inst. for Signal and Information Proc.
and Dept. of Elect. and Comp. Engr.
Location: Mississippi State, MS 39762

4.**Name:** Aravind Ganapathiraju
Affiliation: Inst. for Signal and Information Proc.
and Dept. of Elect. and Comp. Engr.
Location: Mississippi State, MS 39762

5. **Name:** Dr. Joseph Picone
Affiliation: Inst. for Signal and Information Proc.
and Dept. of Elect. and Comp. Engr.
Location: Mississippi State, MS 39762

Author to whom all correspondence should be sent:
Name: Audrey Le
Address: Mississippi State University
Dept. of Elect. and Comp. Engr.
P.O. BOX 9571
Mississippi State, MS 39762
Telephone: (601) 325-8335
Fax: (601) 325-3419
Email: le@isip.msstate.edu

Estimate of paper length: 2 pages
Topic of paper: A Comparison of Classification Algorithms
Key Words: Decision Trees, Classification Trees, Decision Analysis

Checklist for this Manuscript Submission

1. This Manuscript Cover Sheet, completed
2. Four (4) copies of the manuscript included

For Technical Committee's Use

Date Received:
Author Notification Date:
Decision:
Notes:

Manuscript Number Assigned:
Date Camera Ready Received:
Session:

A COMPARISON OF CLASSIFICATION ALGORITHMS ON HIGHLY NONLINEAR SIGNAL PROCESSING PROBLEMS

A. Le, J. Ngan, J. Shaffer, A. Ganapathiraju, J. Picone
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, Mississippi 39762, USA
Ph (601) 325-8335 - Fax (601) 325-3149
{le, ngan, shaffer, ganapath, picone}@isip.msstate.edu

ABSTRACT

A decision tree approach is used to demonstrate high performance classification on two fundamentally different problems: scenic beauty estimates of forestry images and pronunciations of proper nouns. The evaluation databases are comprehensive, including 638 forestry images and over 20,000 surname pronunciations. Decision trees have been constructed using three decision tree algorithms: Bayes, C4, and CART. The preliminary performance of these trees was shown to be promising as those of several standard classification approaches including neural networks, principle components analysis, and linear discriminant analysis. The decision tree technology presented in this paper, which we believe is a significant improvement over existing public domain packages, is freely available from the ISIP web site.

SUMMARY

Decision trees are used in many disciplines and in various applications for data exploration and data classification. Recently, decision trees have enjoyed widespread use on speech and image processing problems. The versatility and usefulness of decision trees in various disciplines motivates us to investigate the functionality of decision trees for two extremely difficult applications — scenic beauty estimates of forestry images and pronunciations of proper nouns. These particular applications, which we have been working on for several years, present highly nonlinear decision spaces with complex interrelationships amongst the data. Conventional technology, such as neural networks, are unable to deliver high performance on these applications.

A decision tree is a data-driven statistical clustering method. As such, the principle drawback of this approach is the need for large amounts of training data representative of the problems. Such statistical approaches tend not to generalize well. The uncertainty in making a decision arises from the fact that the problem is not deterministic but probabilistic and involves emulating human subjective performance. We have little information about some aspects of the problem. In the case of forestry imaging, the system attempts to estimate the aesthetic quality of an image. There is little or no quantitative information on what constitutes a visual pleasing scene — we simply have the output from an extensive human evaluation available as training data. Similarly, predicting proper noun pronunciations is complicated since pronunciations of proper nouns do not follow typical letter-to-sound conversion rules.

We have implemented three types of decision trees: Bayes, C4, and CART. Our Bayesian classifier is based on Bayes' rule. The apriori probabilities are assumed to be known. A input token can be assigned to a class based on a probabilistic decision rule. In our implementation of the Bayesian classifier, we use Bayes splitting rule to build multiple trees and use smoothing to average the trees.

Our implementation of CART, introduced by Brieman, constructs a binary decision tree by recursively partitioning the training data. It grows a large tree to cover all of the training cases and then prunes down the tree to balance the error rate with size of the tree. CART uses the twoing criterion for splitting and cost-complexity cross-validation for pruning. Finally, our C4 tree, introduced by Quinlan, generates a decision tree using the gain ratio as a splitting rule and pessimistic pruning as the pruning rule.

Our baseline technology for providing a scenic beauty estimate of a forestry image involves formulating the problem in a pattern matching paradigm, and to build models based on a diverse mixture of features. The features we use, such as colors, edges, and texture, are extracted from the image and statistically normalized using principle component analysis. We have compared this approach to one in which the features are combined using a decision tree. Table 1 indicates the preliminary results:

Application	Random Guess	Best Non-DT System	Preliminary DT System
Forestry Imaging	67%	37%	47%
Surnames	N/A	50%	In Progress

Table 1. A comparison of classification error rates for two applications: scenic beauty estimates of forestry images and surname pronunciations.

We have yet to evaluate the decision tree trained to generate pronunciations of surnames (used in a speech recognition application). This is a much harder problem due to the complexity of letter-to-sound rules, and the relatively large size of the database.

Our decision tree technology was implemented in C++ and follows a strict object-oriented design methodology. Two major advantages of our software over existing decision tree packages are that it can handle large amounts of training data, and there is no limit on the number of classes, attributes, or attributes values. Our software also allows data tagging enabling each attribute to be selected from the attribute file without having to reformat the training data.