**Contest Instructions**

The objective of this contest is for teams of students to develop the best model to classify forest cover types.

The dataset for the contest contains cartographic variables for forest types in the Roosevelt National Forest of northern Colorado.

More information about the dataset can be found in the metadata file (covertype\_metadata.txt).

Given the cartographic variables, your task is to come up with the **most accurate model to classify the variable cover\_type**.

Please make sure that you read the metadata carefully for information that might help you to create the most accurate classification model.

Remember to explore the dataset before setting out to classify and also keep in mind that **all variables are not needed to create an accurate model**.

In this contest students should use the training data to build and test their model.

The testing data is to be used for prediction and does not have a known class attribute.

Students will submit the classification result from running the model on their testing data.

Submissions are to be submitted to the Classification Contest Blackboard assignment .

**Submissions should contain two files. The first is a csv file containing the ID field and the predicted cover\_type.**

**The second file is a short description of how your set up your classification model including steps for training and testing, the classifier that was used, and any parameters that were set for the classifier.**

Test: 100,001 instances

Train: 481,013 instances

Results:

Basic (Stock from Assignment 2):

Metadata File:

The Forest CoverType dataset

1. Title of Database:

Forest Covertype data

2. Sources:

(a) Original owners of database:

Remote Sensing and GIS Program

Department of Forest Sciences

College of Natural Resources

Colorado State University

Fort Collins, CO 80523

(contact Jock A. Blackard, jblackard 'at' fs.fed.us

or Dr. Denis J. Dean, denis.dean 'at' utdallas.edu)

NOTE: Reuse of this database is unlimited with retention of

copyright notice for Jock A. Blackard and Colorado

State University.

(b) Donors of database:

Jock A. Blackard (jblackard 'at' fs.fed.us)

GIS Coordinator

USFS - Forest Inventory & Analysis

Rocky Mountain Research Station

507 25th Street

Ogden, UT 84401

Dr. Denis J. Dean (denis.dean 'at' utdallas.edu)

Professor

Program in Geography and Geospatial Sciences

School of Economic, Political and Policy Sciences

800 West Campbell Rd

Richardson, TX 75080-3021

Dr. Charles W. Anderson (anderson 'at' cs.colostate.edu)

Associate Professor

Department of Computer Science

Colorado State University

Fort Collins, CO 80523 USA

(c) Date donated: August 1998

3. Past Usage:

Blackard, Jock A. and Denis J. Dean. 2000. "Comparative

Accuracies of Artificial Neural Networks and Discriminant

Analysis in Predicting Forest Cover Types from Cartographic

Variables." Computers and Electronics in Agriculture

24(3):131-151.

Blackard, Jock A. and Denis J. Dean. 1998. "Comparative

Accuracies of Neural Networks and Discriminant Analysis

in Predicting Forest Cover Types from Cartographic

Variables." Second Southern Forestry GIS Conference.

University of Georgia. Athens, GA. Pages 189-199.

Blackard, Jock A. 1998. "Comparison of Neural Networks and

Discriminant Analysis in Predicting Forest Cover Types."

Ph.D. dissertation. Department of Forest Sciences.

Colorado State University. Fort Collins, Colorado.

165 pages.

Abstract of dissertation:

Natural resource managers responsible for developing

ecosystem management strategies require basic descriptive

information including inventory data for forested lands to

support their decision-making processes. However, managers

generally do not have this type of data for inholdings or

neighboring lands that are outside their immediate

jurisdiction. One method of obtaining this information is

through the use of predictive models.

Two predictive models were examined in this study, a

feedforward neural network model and a more traditional

statistical model based on discriminant analysis. The overall

objectives of this research were to first construct these two

predictive models, and second to compare and evaluate their

respective classification accuracies when predicting forest

cover types in undisturbed forests.

The study area included four wilderness areas found in

the Roosevelt National Forest of northern Colorado. A total

of twelve cartographic measures were utilized as independent

variables in the predictive models, while seven major forest

cover types were used as dependent variables. Several subsets

of these variables were examined to determine the best overall

predictive model.

For each subset of cartographic variables examined in

this study, relative classification accuracies indicate the

neural network approach outperformed the traditional

discriminant analysis method in predicting forest cover types.

The final neural network model had a higher absolute

classification accuracy (70.58%) than the final corresponding

linear discriminant analysis model(58.38%). In support of these

classification results, thirty additional networks with randomly

selected initial weights were derived. From these networks, the

overall mean absolute classification accuracy for the neural

network method was 70.52%, with a 95% confidence interval of

70.26% to 70.80%. Consequently, natural resource managers may

utilize an alternative method of predicting forest cover types

that is both superior to the traditional statistical methods and

adequate to support their decision-making processes for

developing ecosystem management strategies.

-- Classification performance

-- first 11,340 records used for training data subset

-- next 3,780 records used for validation data subset

-- last 565,892 records used for testing data subset

-- 70% Neural Network (backpropagation)

-- 58% Linear Discriminant Analysis

4. Relevant Information Paragraph:

Predicting forest cover type from cartographic variables only

(no remotely sensed data). The actual forest cover type for

a given observation (30 x 30 meter cell) was determined from

US Forest Service (USFS) Region 2 Resource Information System

(RIS) data. Independent variables were derived from data

originally obtained from US Geological Survey (USGS) and

USFS data. Data is in raw form (not scaled) and contains

binary (0 or 1) columns of data for qualitative independent

variables (wilderness areas and soil types).

This study area includes four wilderness areas located in the

Roosevelt National Forest of northern Colorado. These areas

represent forests with minimal human-caused disturbances,

so that existing forest cover types are more a result of

ecological processes rather than forest management practices.

Some background information for these four wilderness areas:

Neota (area 2) probably has the highest mean elevational value of

the 4 wilderness areas. Rawah (area 1) and Comanche Peak (area 3)

would have a lower mean elevational value, while Cache la Poudre

(area 4) would have the lowest mean elevational value.

As for primary major tree species in these areas, Neota would have

spruce/fir (type 1), while Rawah and Comanche Peak would probably

have lodgepole pine (type 2) as their primary species, followed by

spruce/fir and aspen (type 5). Cache la Poudre would tend to have

Ponderosa pine (type 3), Douglas-fir (type 6), and

cottonwood/willow (type 4).

The Rawah and Comanche Peak areas would tend to be more typical of

the overall dataset than either the Neota or Cache la Poudre, due

to their assortment of tree species and range of predictive

variable values (elevation, etc.) Cache la Poudre would probably

be more unique than the others, due to its relatively low

elevation range and species composition.

5. Number of instances (observations): 581,012

6. Number of Attributes: 12 measures, but 54 columns of data

(10 quantitative variables, 4 binary

wilderness areas and 40 binary

soil type variables)

7. Attribute information:

Given is the attribute name, attribute type, the measurement unit and

a brief description. The forest cover type is the classification

problem. The order of this listing corresponds to the order of

numerals along the rows of the database.

Name Data Type Measurement Description

Elevation quantitative meters Elevation in meters

Aspect quantitative azimuth Aspect in degrees azimuth

Slope quantitative degrees Slope in degrees

Horizontal\_Distance\_To\_Hydrology quantitative meters Horz Dist to nearest surface water features

Vertical\_Distance\_To\_Hydrology quantitative meters Vert Dist to nearest surface water features

Horizontal\_Distance\_To\_Roadways quantitative meters Horz Dist to nearest roadway

Hillshade\_9am quantitative 0 to 255 index Hillshade index at 9am, summer solstice

Hillshade\_Noon quantitative 0 to 255 index Hillshade index at noon, summer soltice

Hillshade\_3pm quantitative 0 to 255 index Hillshade index at 3pm, summer solstice

Horizontal\_Distance\_To\_Fire\_Points quantitative meters Horz Dist to nearest wildfire ignition points

Wilderness\_Area (4 binary columns) qualitative 0 (absence) or 1 (presence) Wilderness area designation

Soil\_Type (**40** binary columns) qualitative 0 (absence) or 1 (presence) Soil Type designation

**Cover\_Type (7 types) integer 1 to 7 Forest Cover Type designation**

Code Designations:

Wilderness Areas: 1 -- Rawah Wilderness Area

2 -- Neota Wilderness Area

3 -- Comanche Peak Wilderness Area

4 -- Cache la Poudre Wilderness Area

Soil Types: 1 to 40 : based on the USFS Ecological

Landtype Units (ELUs) for this study area:

Study Code USFS ELU Code Description

1 2702 Cathedral family - Rock outcrop complex, extremely stony.

2 2703 Vanet - Ratake families complex, very stony.

3 2704 Haploborolis - Rock outcrop complex, rubbly.

4 2705 Ratake family - Rock outcrop complex, rubbly.

5 2706 Vanet family - Rock outcrop complex complex, rubbly.

6 2717 Vanet - Wetmore families - Rock outcrop complex, stony.

7 3501 Gothic family.

8 3502 Supervisor - Limber families complex.

9 4201 Troutville family, very stony.

10 4703 Bullwark - Catamount families - Rock outcrop complex, rubbly.

11 4704 Bullwark - Catamount families - Rock land complex, rubbly.

12 4744 Legault family - Rock land complex, stony.

13 4758 Catamount family - Rock land - Bullwark family complex, rubbly.

14 5101 Pachic Argiborolis - Aquolis complex.

15 5151 unspecified in the USFS Soil and ELU Survey.

16 6101 Cryaquolis - Cryoborolis complex.

17 6102 Gateview family - Cryaquolis complex.

18 6731 Rogert family, very stony.

19 7101 Typic Cryaquolis - Borohemists complex.

20 7102 Typic Cryaquepts - Typic Cryaquolls complex.

21 7103 Typic Cryaquolls - Leighcan family, till substratum complex.

22 7201 Leighcan family, till substratum, extremely bouldery.

23 7202 Leighcan family, till substratum - Typic Cryaquolls complex.

24 7700 Leighcan family, extremely stony.

25 7701 Leighcan family, warm, extremely stony.

26 7702 Granile - Catamount families complex, very stony.

27 7709 Leighcan family, warm - Rock outcrop complex, extremely stony.

28 7710 Leighcan family - Rock outcrop complex, extremely stony.

29 7745 Como - Legault families complex, extremely stony.

30 7746 Como family - Rock land - Legault family complex, extremely stony.

31 7755 Leighcan - Catamount families complex, extremely stony.

32 7756 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.

33 7757 Leighcan - Catamount families - Rock outcrop complex, extremely stony.

34 7790 Cryorthents - Rock land complex, extremely stony.

35 8703 Cryumbrepts - Rock outcrop - Cryaquepts complex.

36 8707 Bross family - Rock land - Cryumbrepts complex, extremely stony.

37 8708 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.

38 8771 Leighcan - Moran families - Cryaquolls complex, extremely stony.

39 8772 Moran family - Cryorthents - Leighcan family complex, extremely stony.

40 8776 Moran family - Cryorthents - Rock land complex, extremely stony.

Note: First digit: climatic zone Second digit: geologic zones

1. lower montane dry 1. alluvium

2. lower montane 2. glacial

3. montane dry 3. shale

4. montane 4. sandstone

5. montane dry and montane 5. mixed sedimentary

6. montane and subalpine 6. unspecified in the USFS ELU Survey

7. subalpine 7. igneous and metamorphic

8. alpine 8. volcanic

The third and fourth ELU digits are unique to the mapping unit

and have no special meaning to the climatic or geologic zones.

Forest Cover Type Classes: 1 -- Spruce/Fir

2 -- Lodgepole Pine

3 -- Ponderosa Pine

4 -- Cottonwood/Willow

5 -- Aspen

6 -- Douglas-fir

7 -- Krummholz

8. Basic Summary Statistics for quantitative variables only

(whole dataset -- thanks to Phil Rennert for the summary values):

Name Units Mean Std Dev

Elevation meters 2959.36 279.98

Aspect azimuth 155.65 111.91

Slope degrees 14.10 7.49

Horizontal\_Distance\_To\_Hydrology meters 269.43 212.55

Vertical\_Distance\_To\_Hydrology meters 46.42 58.30

Horizontal\_Distance\_To\_Roadways meters 2350.15 1559.25

Hillshade\_9am 0 to 255 index 212.15 26.77

Hillshade\_Noon 0 to 255 index 223.32 19.77

Hillshade\_3pm 0 to 255 index 142.53 38.27

Horizontal\_Distance\_To\_Fire\_Points meters 1980.29 1324.19

9. Missing Attribute Values: None.

10. Class distribution:

Number of records of Spruce-Fir: 211840

Number of records of Lodgepole Pine: 283301

Number of records of Ponderosa Pine: 35754

Number of records of Cottonwood/Willow: 2747

Number of records of Aspen: 9493

Number of records of Douglas-fir: 17367

Number of records of Krummholz: 20510

Number of records of other: 0

Total records: 581012

=====================================================================

Jock A. Blackard

08/28/1998 -- original text

12/07/1999 -- updated mailing address, citations, background info

for study area, added summary statistics.

=====================================================================