ADCG WS 2016/17 Challenge 01: Satellite Image Data Set.

FILE NAMES

sat-train.csv.dat – training set, label in the last column
sat-test-data.csv.dat - test set in csv format

PURPOSE

The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. The aim is to use supervised anomaly detection in order to detect cotton soil.

PROBLEM TYPE

Binary Classification

SOURCE

The small sample database was provided by:
Ashwin Srinivasan
Department of Statistics and Modelling Science
University of Strathclyde
Glasgow
Scotland
UK

ORIGIN

The original Landsat data for this database was generated from data purchased from NASA by the Australian Centre for Remote Sensing, and used for research at:

The Centre for Remote Sensing
University of New South Wales
Kensington, PO Box 1
NSW 2033
Australia.

The sample database was generated taking a small section (82 rows and 100 columns) from the original data. The binary values were converted to their present ASCII form by Ashwin Srinivasan. The classification for each pixel was performed on the basis of an actual site visit by Ms. Karen Hall, when working for Professor John A. Richards, at the Centre for Remote Sensing at the University of New South Wales, Australia. Conversion to 3x3 neighbourhoods and splitting into test and training sets was done by Alistair Sutherland.

DESCRIPTION

One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340 x 3380 such pixels.

The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighbourhood of pixels completely contained within the 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighbourhood and a number indicating the classification label of the central pixel. The number is a code for the following classes:

Number	Class
0	normal soil
1	cotton soil

The data is given in random order and certain lines of data have been removed so you cannot reconstruct the original image from this dataset.

In each line of data the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom.

NUMBER OF EXAMPLES

training set	4435
test set	2000

NUMBER OF ATTRIBUTES

36 (= 4 spectral bands x 9 pixels in neighbourhood)

ATTRIBUTES

The attributes are numerical, in the range 0 to 255.

CLASS

There are 2 decision classes

MISSING VALUES

The dataset was purposely preprocessed to have missing values on 70% of training instances, each of whose 30% feature values are removed and replaced with a NaN value.

ORIGINAL AUTHOR

Ashwin Srinivasan
Department of Statistics and Data Modeling
University of Strathclyde
Glasgow
Scotland
UK
ross@uk.ac.turing