FILE NAMES

sat.trn - training set

sat.tst - test set

!!! NB. DO NOT USE CROSS-VALIDATION WITH THIS DATASET !!!

Just train and test only once with the above

training and test sets.

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 predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number.

PROBLEM TYPE

Classification

AVAILABLE

This database was generated from Landsat Multi-Spectral Scanner image data. These and other forms of remotely sensed imagery can be purchased at a price from relevant governmental authorities. The data is usually in binary form, and distributed on magnetic tape(s).

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.

HISTORY

The Landsat satellite data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multispectral and radar data, maps indicating topography, land use etc. is expected to assume significant importance with the onset of an era characterised by integrative approaches to remote sensing (for example, NASA's Earth Observing System commencing this decade). Existing statistical methods are ill-equipped for handling such diverse data types. Note that this is not true for Landsat MSS data considered in isolation (as in this sample database). This data satisfies the important requirements of being numerical and at a single resolution, and standard maximum-likelihood classification performs very well. Consequently, for this data, it should be interesting to compare the performance of other methods against the statistical approach.

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

1 red soil

2 cotton crop

3 grey soil

4 damp grey soil

5 soil with vegetation stubble

6 mixture class (all types present)

7 very damp grey soil

NB. There are no examples with class 6 in this dataset.

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. Thus, the four spectral values for the central pixel are given by attributes 17,18,19 and 20. If you like you can use only these four attributes, while ignoring the others. This avoids the problem which arises when a 3x3 neighbourhood straddles a boundary.

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 6 decision classes: 1,2,3,4,5 and 7.

NB. There are no examples with class 6 in this dataset- they have all been removed because of doubts about the validity of this class.

AUTHOR

Ashwin Srinivasan

Department of Statistics and Data Modeling

University of Strathclyde

Glasgow

Scotland

UK

ross@uk.ac.turing

TÊN TỆP TIN sat.trn - tập huấn luyện sat.tst - tập kiểm tra

Copy code

!!! LƯU Ý. KHÔNG SỬ DỤNG KIỂM CHỨNG CHÉO VỚI TẬP DỮ LIỆU NÀY !!!

Chỉ huấn luyện và kiểm tra một lần với

các tập huấn luyện và kiểm tra ở trên

MỤC ĐÍCH Cơ sở dữ liệu bao gồm các giá trị phổ đa phổ của các điểm ảnh trong các lân cận 3x3 trong một ảnh vệ tinh, và phân loại liên kết với điểm ảnh trung tâm trong mỗi lân cận. Mục tiêu là dự đoán phân loại này, dựa trên các giá trị phổ đa phổ. Trong cơ sở dữ liệu mẫu, lớp của một điểm ảnh được mã hóa là một số.

LOẠI VẤN ĐỀ Phân loại

KHẢ DỤNG Cơ sở dữ liệu này được tạo ra từ dữ liệu ảnh Máy quét đa phổ Landsat. Những dạng ảnh viễn thám này và các dạng khác có thể mua với giá từ các cơ quan chính phủ liên quan. Dữ liệu thường ở dạng nhị phân, và phân phối trên băng từ (các băng từ).

NGUỒN Cơ sở dữ liệu mẫu nhỏ được cung cấp bởi: Ashwin Srinivasan Khoa Thống kê và Khoa học Mô hình hóa Đại học Strathclyde Glasgow Scotland UK

NGUỒN GỐC Dữ liệu Landsat gốc cho cơ sở dữ liệu này được tạo ra từ dữ liệu mua từ NASA bởi Trung tâm Viễn thám Úc, và được sử dụng cho nghiên cứu tại: Trung tâm Viễn thám Đại học New South Wales Kensington, PO Box 1 NSW 2033 Australia.

Copy code

Cơ sở dữ liệu mẫu được tạo ra bằng cách lấy một phần nhỏ (82 hàng và 100 cột) từ dữ liệu gốc. Các giá trị nhị phân được chuyển đổi sang dạng ASCII hiện tại bởi Ashwin Srinivasan. Việc phân loại cho mỗi điểm ảnh được thực hiện dựa trên thực tế khảo sát thực địa bởi bà Karen Hall, khi làm việc cho Giáo sư John A. Richards, tại Trung tâm Viễn thám tại Đại học New South Wales, Úc. Việc chuyển đổi thành lân cận 3x3 và chia thành các tập kiểm tra và huấn luyện được thực hiện bởi Alistair Sutherland.

LỊCH SỬ Dữ liệu vệ tinh Landsat là một trong nhiều nguồn thông tin có sẵn cho một cảnh. Việc diễn giải một cảnh bằng cách tích hợp dữ liệu không gian đa dạng kiểu và độ phân giải bao gồm dữ liệu đa phổ và radar, các bản đồ cho biết địa hình, sử dụng đất, v.v. được kỳ vọng sẽ đảm nhận tầm quan trọng đáng kể với sự khởi đầu của một kỷ nguyên đặc trưng bởi các cách tiếp cận tổng hợp đối với viễn thám (ví dụ, Hệ thống Quan sát Trái đất của NASA bắt đầu thập kỷ này). Các phương pháp thống kê hiện tại không được trang bị tốt để xử lý các kiểu dữ liệu đa dạng như vậy. Lưu ý rằng điều này không đúng với dữ liệu Landsat MSS được xem xét riêng lẻ (như trong cơ sở dữ liệu mẫu này). Dữ liệu này đáp ứng các yêu cầu quan trọng là số hóa và ở cùng độ phân giải, và phân loại xác suất cực đại chuẩn hoạt động rất tốt. Do đó, đối với dữ liệu này, việc so sánh hiệu suất của các phương pháp khác với cách tiếp cận thống kê sẽ rất thú vị.

MÔ TẢ Một khung ảnh MSS Landsat bao gồm bốn hình ảnh kỹ thuật số của cùng một cảnh ở các băng phổ khác nhau. Hai trong số đó nằm trong vùng khả kiến (tương ứng xấp xỉ với vùng màu xanh và đỏ của quang phổ khả kiến) và hai ở vùng hồng ngoại gần. Mỗi điểm ảnh là một từ nhị phân 8 bit, với 0 tương ứng với màu đen và 255 tương ứng với màu trắng. Độ phân giải không gian của một điểm ảnh là khoảng 80m x 80m. Mỗi hình ảnh chứa 2340 x 3380 điểm ảnh như vậy.

Copy code

Cơ sở dữ liệu là một (rất nhỏ) tiểu vùng của một cảnh, gồm 82 x 100 điểm ảnh. Mỗi dòng dữ liệu tương ứng với một lân cận vuông 3x3 điểm ảnh nằm hoàn toàn trong tiểu vùng 82x100. Mỗi dòng chứa các giá trị điểm ảnh trong bốn băng phổ (đã chuyển sang ASCII) của mỗi 9 điểm ảnh trong lân cận 3x3 và một số cho biết nhãn phân loại của điểm ảnh trung tâm. Số đó là mã của các lớp sau:

Số Lớp

1 đất đỏ

2 cây bông

3 đất xám

4 đất xám ẩm

5 đất có còn rơm rạ

6 lớp hỗn hợp (tất cả các loại đều có)

7 đất xám rất ẩm

LƯU Ý. Không có các ví dụ với lớp 6 trong tập dữ liệu này.

Dữ liệu được cho theo thứ tự ngẫu nhiên và một số dòng dữ liệu đã bị xóa để bạn không thể tái tạo lại hình ảnh gốc từ tập dữ liệu này.

Trong mỗi dòng dữ liệu, bốn giá trị phổ cho điểm ảnh góc trên bên trái được cho trước tiên, tiếp theo là bốn giá trị phổ cho điểm ảnh giữa phía trên và sau đó là của điểm ảnh phía trên bên phải, và cứ thế với các điểm ảnh được đọc theo trình tự từ trái sang phải và từ trên xuống dưới. Như vậy, bốn giá trị phổ của điểm ảnh trung tâm được cho bởi các thuộc tính 17,18,19 và 20. Nếu mu ốn bạn có thể chỉ sử dụng bốn thuộc tính này, trong khi bỏ qua các thuộc tính khác. Điều này tránh vấn đề phát sinh khi một lân cận 3x3 bao trùm một ranh giới.

SỐ LƯỢNG VÍ DỤ tập huấn luyện 4435 tập kiểm tra 2000

SỐ LƯỢNG THUỘC TÍNH 36 (= 4 băng phổ x 9 điểm ảnh trong lân cận)

THUỘC TÍNH Các thuộc tính là số, trong khoảng 0 đến 255.

LỚP Có 6 lớp quyết định: 1, 2, 3, 4, 5 và 7.

Copy code

LƯU Ý. Không có các ví dụ với lớp 6 trong tập dữ liệu này - tất cả đều đã bị xóa vì nghi ngờ về tính hợp lệ của lớp này.

TÁC GIẢ Ashwin Srinivasan Khoa Thống kê và Mô hình hóa Dữ liệu  
Đại học Strathclyde Glasgow Scotland UK