

**Analysis of satellite images:
Identification of the most relevant parameters
for classification problems**

MIKAN

Dr. Martin Fischer

27 rue de Sébastopol

98800 Nouméa, New Caledonia

Abstract

MARS (Multivariate Adaptive Regression Splines) was used to automatically identify the most relevant parameters for a classification problem. In the present case the algorithm yields that out of 68 parameters only two are needed to separate the sample data into “deciduous trees” and “coniferous trees”. Compared to the number of input parameters the sample size was very small (18), which shows that the algorithm can deal with rather low signal to noise ratios. The computational time (Pentium III, 900 MHz) was less than a second in the present case.

Introduction

The output of an object oriented imaging analysis tool is being analysed. The tool automatically identifies objects on images by evaluating and attributing between 50 and 300 parameters to each object. Such parameters describe features like shape, size, brightness, and colors, among others. Currently the producer of the software is seeking for methods to automatically identify the most relevant parameters for classifying the numerous objects, visible on such images.

In this study we investigate the use of multivariate adaptive regression splines (MARS) for this purpose. Data provided by the producer are used. The data were derived from an aerial photography showing a landscape near Berlin. The software identified 3017 different objects and attributed 74 parameters to each object. In this study we apply MARS to identify the most important parameters to distinguish between deciduous trees and coniferous trees.

Data

Data Description

A data set containing 3017 records was provided by the producer. Each record contains 74 parameters to describe the respective object. For 30 data records (samples) the kind of object, described by the respective record, is specified. The samples contain 6 different object types. The types, and the number of samples given for each type are listed in the table below:

type	# of samples
coniferous trees	8
coniferous shadows	4
deciduous shadows	4
deciduous trees	10
grass	2
non vegetation	1

Considering the number of parameters used to describe each object (74), the number of samples is rather small for each sample type. In the case of “non vegetation” and “grass” it is definitely insufficient for any statistical method, and also for “deciduous shadows” and “coniferous shadows” a reasonable result is at least questionable. In this study we therefore focus on two types only, namely “deciduous trees” and “coniferous trees”.

Data Processing

Some of the records in the sample data set contain missing values. The corresponding attributes (parameters) cannot be used for classification (not available for all data sets) and are therefore removed from the analyses. Out of the 74 parameters in the original data set 68 were used, which are listed below.

- [1] "Brightness"
- [2] "Mean.3697.5.tif..1."
- [3] "Stdev.3697.5.tif..1."
- [4] "Ratio.3697.5.tif..1."
- [5] "Mean.diff..to.neighbors.3697.5.tif..1....0."
- [6] "Mean.diff..to.neighbors..abs..3697.5.tif..1....0."
- [7] "Rel..border.to.brighter.neighbors.3697.5.tif..1."
- [8] "Mean.diff..to.scene.3697.5.tif..1."
- [9] "Mean.of.sub.objects..stddev.3697.5.tif..1....1."
- [10] "Avrg..mean.diff.to.neighbors.of.sub.objects.3697.5.tif..1....1."
- [11] "Mean.3697.5.tif..2."
- [12] "Stdev.3697.5.tif..2."
- [13] "Ratio.3697.5.tif..2."
- [14] "Mean.diff..to.neighbors.3697.5.tif..2....0."
- [15] "Mean.diff..to.neighbors..abs..3697.5.tif..2....0."
- [16] "Rel..border.to.brighter.neighbors.3697.5.tif..2."
- [17] "Mean.diff..to.scene.3697.5.tif..2."
- [18] "Mean.of.sub.objects..stddev.3697.5.tif..2....1."
- [19] "Avrg..mean.diff.to.neighbors.of.sub.objects.3697.5.tif..2....1."
- [20] "Mean.3697.5.tif..3."
- [21] "Stdev.3697.5.tif..3."
- [22] "Ratio.3697.5.tif..3."
- [23] "Mean.diff..to.neighbors.3697.5.tif..3....0."
- [24] "Mean.diff..to.neighbors..abs..3697.5.tif..3....0."
- [25] "Rel..border.to.brighter.neighbors.3697.5.tif..3."
- [26] "Mean.diff..to.scene.3697.5.tif..3."
- [27] "Mean.of.sub.objects..stddev.3697.5.tif..3....1."
- [28] "Avrg..mean.diff.to.neighbors.of.sub.objects.3697.5.tif..3....1."
- [29] "Area"
- [30] "Length"
- [31] "Width"
- [32] "Border.length"
- [33] "Length.width"
- [34] "Shape.index"
- [35] "Density"
- [36] "Main.direction"
- [37] "Asymmetry"
- [38] "Area..excluding.inner.polygons."
- [39] "Area..including.inner.polygons."
- [40] "Perimeter..polygon."
- [41] "Compactness..polygon."
- [42] "Number.of.edges..polygon."
- [43] "Stddev.of.length.of.edges..polygon."
- [44] "Average.length.of.edges..polygon."
- [45] "Length.of.longest.edge..polygon."
- [46] "Number.of.inner.objects..polygon."
- [47] "Edges.longer.than..polygon....10."
- [48] "Rectangular.angles.with.edges.longer.than..polygon....10."
- [49] "Area.of.sub.objects..mean...1."
- [50] "Area.of.sub.objects..stddev...1."
- [51] "Density.of.sub.objects..mean...1."
- [52] "Density.of.sub.objects..stddev...1."
- [53] "Asymmetry.of.sub.objects..mean...1."
- [54] "Asymmetry.of.sub.objects..stddev...1."
- [55] "Direction.of.sub.objects..mean...1."
- [56] "Direction.of.sub.objects..stddev...1."

[57] "Mean.diff..to.neighbors.3697.5.tif..1....10."
 [58] "Mean.diff..to.neighbors.3697.5.tif..2....10."
 [59] "Mean.diff..to.neighbors.3697.5.tif..3....10."
 [60] "Mean.diff..to.neighbors.3697.5.tif..1....20."
 [61] "Mean.diff..to.neighbors.3697.5.tif..2....20."
 [62] "Mean.diff..to.neighbors.3697.5.tif..3....20."
 [63] "Mean.diff..to.neighbors..abs..3697.5.tif..1....10."
 [64] "Mean.diff..to.neighbors..abs..3697.5.tif..2....10."
 [65] "Mean.diff..to.neighbors..abs..3697.5.tif..3....10."
 [66] "Mean.diff..to.neighbors..abs..3697.5.tif..1....20."
 [67] "Mean.diff..to.neighbors..abs..3697.5.tif..2....20."
 [68] "Mean.diff..to.neighbors..abs..3697.5.tif..3....20."

To ensure that all parameters are equally weighted during the analysis they are normalized to have variance equal to one. To compute the normalization factors all 3017 data sets are used.

The analyses in the following section are carried out with a sample data set of 18 members (10 deciduous trees and 8 coniferous trees) and 68 attributes for each member.

Analyses

The goal of this study is to automatically identify out of 68 parameters, those that are most relevant to distinguish between deciduous trees and coniferous trees. For this purpose MARS was applied as a classification tool to the sample data. MARS is a statistical modeling method that balances complexity of the final model against accuracy on the training data set. The algorithm uses the so called "generalized cross validation" (gcv) concept for optimal model selection. The value of "gcv" can be specified as an input parameter to the MARS algorithm. The residual sum of squares of a model is penalized by dividing by the square of $[1-(gcv * \text{model size})/\text{cases}]$. A larger gcv value would tend to produce a smaller model. As a result MARS provides a list of relevant input variable (predictors).

Specifying a relatively high gcv value forces MARS to search for very simple models. This corresponds to the problem of finding the smallest possible number of predictors that are needed for a classification problem. We use two different values for gcv to find models of different complexity:

experiment	gcv
Exp01	4
Exp02	2

In both experiments the maximum allowed interaction degree between input variables is set to 2. This means that MARS builds models out of basis functions that contain one or two input variables (e.g. $a*\text{var1} + b*\text{var1}*\text{var2} + c*\text{var1}*\text{var3} + \dots$).

As a reference we recapitulate some results from a detailed study made by the producer.

Study carried out by the producer

In this (lengthy) study it was found that variable 4 ("Ratio.3697.5.tif..1.") is the most important parameter to distinguish between the two types of trees. However, using only this parameter is not sufficient for a perfect separation between the two types of trees. At least one sample member is misclassified if only this parameter is used (compare Fig.2). To reduce the number of wrongly classified sample members, variable 1 ("brightness") was added as a second criterion for classification. Now the tree types can be separated by a single line (Fig. 1). However, the distance between the two groups is rather small.

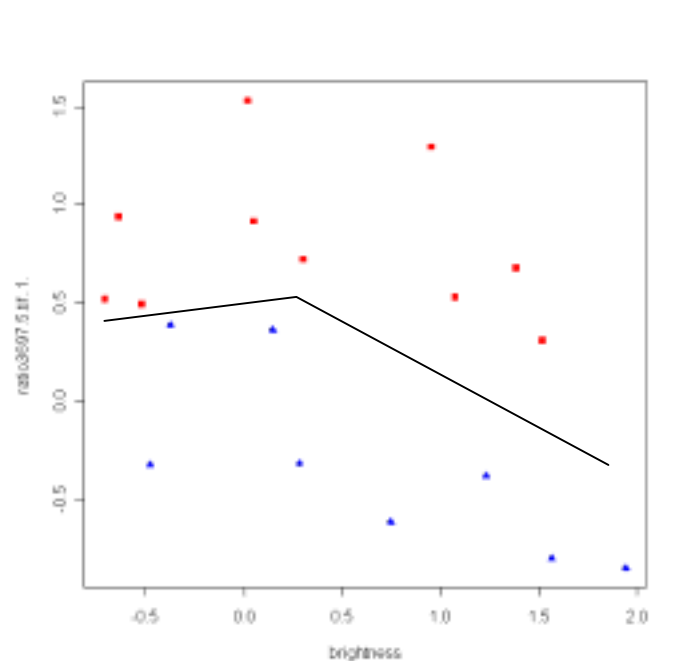


Figure 1: Deciduous trees (red squares) and coniferous trees (blue triangles) shown in a two-dimensional diagram using brightness (variable 1) and ratio3697.5.tif..1. (variable 4).

Exp01

MARS was run with a gcv value of 4. This is a relatively large value, which therefore forces the algorithm to search for a very simple model (very few model components).

The final model is a function of variable 4 ("Ratio.3697.5.tif..1.") only.

Displaying the sample data in a one-dimensional graph with variable 4 on the vertical axis shows that a horizontal line at $y=0.45$ separates the two types almost perfectly. This result corresponds to the findings from the detailed data analysis made by the producer. An alternative threshold to separate the two groups could be at $y=0$. In that case the distance between the two groups is bigger, but two sample members are misclassified (compare Fig. 2).

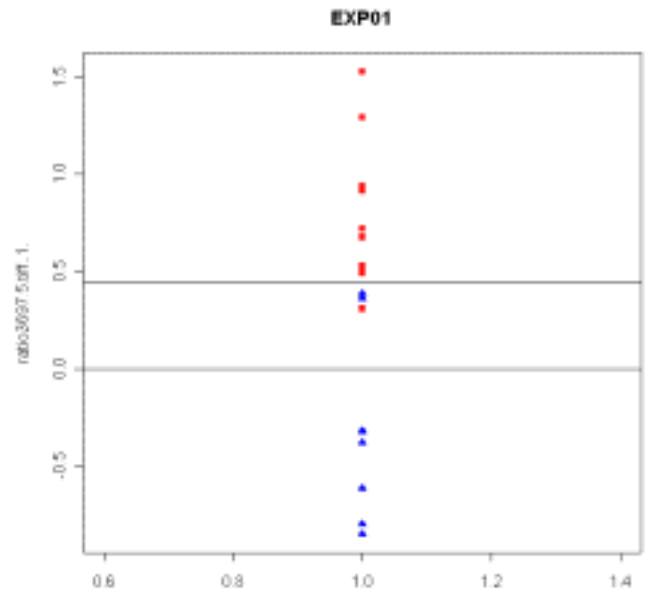


Figure 2: Sample data displayed in a one-dimensional diagram, using variable 4 (ratio3697.5.tif..1.) only. The horizontal lines show two possible thresholds to separate the two groups.

Exp02

The setup is exactly the same as in Exp01, only the gcv value is reduced to 2. In this case MARS becomes more sensitive to small scale features, and the model becomes more complex. It is now formed of two basis functions. Variable 4 remains the first – and most important – term. The second basis function depends linearly on variable 66 (“Mean.diff..to.neighbors..abs..3697.5.tif..1....20.”). Displaying the sample data in a two-dimensional graph (Fig. 3) with variable 66 along the horizontal axis and variable 4 along the vertical axis reveals that these two parameters enable a perfect separation between the two groups (black line). The shape of the separation line is simpler than in Fig. 1 and the distance between the two groups is also bigger. This shows that the combination of variable 4 and variable 66 is a better choice to separate the sample data than variable 4 and variable 1.

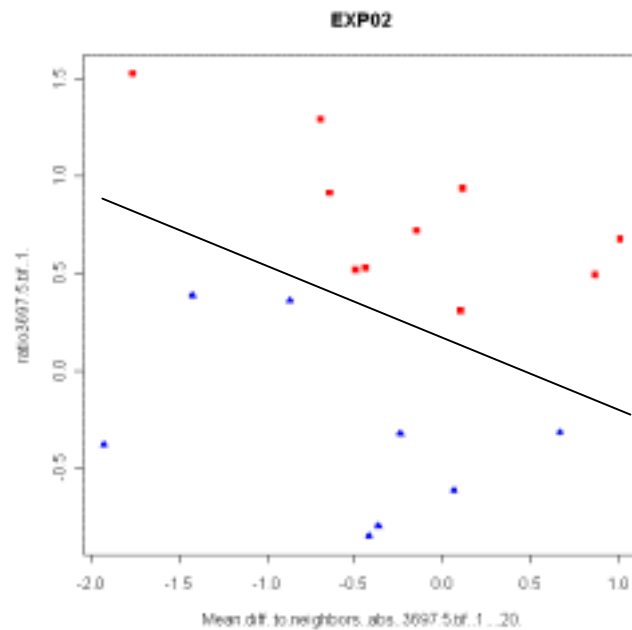


Figure 3: Sample data shown in a two-dimensional diagram using variable 66 (Mean.dif..to.neighbors..abs..3697.5.tif..1....20.) and variable 4 (ratio3697.5.tif..1.). The black line indicates a possible way to separate the two groups.

Summary

The results of the analyses can be summarized as follows:

1. MARS automatically identified variable 4 as the most important variable for the classification problem. No manual interfering or special data preprocessing was necessary. The result agrees with the findings from a manual study made by the producer.
2. This was achieved despite the fact that the number of variables (68) is almost four times larger than the sample size (18).
3. With a reduced gcv MARS identifies a second variable which allows an even better separation between the two groups.
4. In the present case the required CPU time was less than a second on a 900 MHz Pentium III processor.

Conclusions and Outlook

Out of 68 parameters, MARS found the most relevant variables to separate the sample data into “deciduous trees” and “coniferous trees”. The result was achieved fully automatically and within less than a second. From this we infer that MARS is a powerful tool to identify the most relevant parameters for classification problems automatically.

In the present study the response variable had only two valid states, namely “deciduous trees” and “coniferous trees”. In practical applications, however, the response variable often has more than two levels. Although MARS can cope with multivariate output variables, problems occur when the sample size – as in the present case – is very small. Therefore strategies must be developed to treat multivariate response variables in the presence of small sample sizes. A possible solution would be to do the separation mutually for

each possible pair of response levels, or to reduce the number of input variables in a preprocessing step.