1. Title of Database: Blocks Classification

2. Sources:

(a) Donato Malerba

Dipartimento di Informatica

University of Bari

via Orabona 4

70126 Bari - Italy

phone: +39 - 80 - 5443269

fax: +39 - 80 - 5443196

malerbad@vm.csata.it

(b) Donor: Donato Malerba

(c) Date: July 1995

3. Past Usage:

This data set have been used to try different simplification methods

for decision trees. A summary of the results can be found in:

Malerba, D., Esposito, F., and Semeraro, G.

"A Further Comparison of Simplification Methods for Decision-Tree Induction."

In D. Fisher and H. Lenz (Eds.), "Learning from Data:

Artificial Intelligence and Statistics V", Lecture Notes in Statistics,

Springer Verlag, Berlin, 1995.

The problem consists in classifying all the blocks of the page

layout of a document that has been detected by a segmentation

process. This is an essential step in document analysis

in order to separate text from graphic areas. Indeed,

the five classes are: text (1), horizontal line (2),

picture (3), vertical line (4) and graphic (5).

For a detailed presentation of the problem see:

Esposito F., Malerba D., & Semeraro G.

Multistrategy Learning for Document Recognition

Applied Artificial Intelligence, 8, pp. 33-84, 1994

All instances have been personally checked so that

low noise is present in the data.

4. Relevant Information Paragraph:

The 5473 examples comes from 54 distinct documents.

Each observation concerns one block.

All attributes are numeric.

Data are in a format readable by C4.5.

5. Number of Instances: 5473.

6. Number of Attributes

height: integer. | Height of the block.

lenght: integer. | Length of the block.

area: integer. | Area of the block (height \* lenght);

eccen: continuous. | Eccentricity of the block (lenght / height);

p\_black: continuous. | Percentage of black pixels within the block (blackpix / area);

p\_and: continuous. | Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area);

mean\_tr: continuous. | Mean number of white-black transitions (blackpix / wb\_trans);

blackpix: integer. | Total number of black pixels in the original bitmap of the block.

blackand: integer. | Total number of black pixels in the bitmap of the block after the RLSA.

wb\_trans: integer. | Number of white-black transitions in the original bitmap of the block.

7. Missing Attribute Values: No missing value.

8. Class Distribution:

Valid Cum

Class Frequency Percent Percent Percent

text 4913 89.8 89.8 89.8

horiz. line 329 6.0 6.0 95.8

graphic 28 .5 .5 96.3

vert. line 88 1.6 1.6 97.9

picture 115 2.1 2.1 100.0

------- ------- -------

TOTAL 5473 100.0 100.0

Summary Statistics:

Variable Mean Std Dev Minimum Maximum Correlation

HEIGHT 10.47 18.96 1 804 .3510

LENGTH 89.57 114.72 1 553 -.0045

AREA 1198.41 4849.38 7 143993 .2343

ECCEN 13.75 30.70 .007 537.00 .0992

P\_BLACK .37 .18 .052 1.00 .2130

P\_AND .79 .17 .062 1.00 -.1771

MEAN\_TR 6.22 69.08 1.00 4955.00 .0723

BLACKPIX 365.93 1270.33 7 33017 .1656

BLACKAND 741.11 1881.50 7 46133 .1565

WB\_TRANS 106.66 167.31 1 3212 .0337