1. Title of Database: Internet advertisements

2. Sources:

(a) Creator & donor: Nicholas Kushmerick <nick@ucd.ie>

(c) Generated: April-July 1998

3. Past Usage:

N. Kushmerick (1999). "Learning to remove Internet advertisements",

3rd Int Conf Autonomous Agents. Available at

www.cs.ucd.ie/staff/nick/research/download/kushmerick-aa99.ps.gz.

Accuracy >97% using C4.5rules in predicting whether an image is an

advertisement.

4. This dataset represents a set of possible advertisements on

Internet pages. The features encode the geometry of the image (if

available) as well as phrases occuring in the URL, the image's URL and

alt text, the anchor text, and words occuring near the anchor text.

The task is to predict whether an image is an advertisement ("ad") or

not ("nonad").

5. Number of Instances: 3279 (2821 nonads, 458 ads)

6. Number of Attributes: 1558 (3 continous; others binary; this is the

"STANDARD encoding" mentioned in the [Kushmerick, 99].)

One or more of the three continous features are missing in 28%

of the instances; missing values should be interpreted as "unknown".

7. See [Kushmerick, 99] for details of the attributes; in

".names" format:

height: continuous. | possibly missing

width: continuous. | possibly missing

aratio: continuous. | possibly missing

local: 0,1.

| 457 features from url terms, each of the form "url\*term1+term2...";

| for example:

url\*images+buttons: 0,1.

...

| 495 features from origurl terms, in same form; for example:

origurl\*labyrinth: 0,1.

...

| 472 features from ancurl terms, in same form; for example:

ancurl\*search+direct: 0,1.

...

| 111 features from alt terms, in same form; for example:

alt\*your: 0,1.

...

| 19 features from caption terms

caption\*and: 0,1.

...

8. Missing Attribute Values: how many per each attribute?

28% of instances are missing some of the continous attributes.

9. Class Distribution: number of instances per class

2821 nonads, 458 ads.