**95.Farm Ads**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Farm+Ads

2. 数据库描述

【1.[数据集名称]数据集由[机构名或人名]采集；】The data used in our experiments were collected by E. Alpaydin, C. Kaynak, from Department of Computer Engineering,Bogazici University at July,1998.【2.用于[什么实验目的]】We used preprocessing programs made available by NIST to extract normalized bitmaps of handwritten digits from a preprinted form.【3】

【4】The database has 5620 samples, respectively belong to optdigits.tra with 3823 samples and optidigits.tes with 1797 samples. The categories of network system include seven categories, as shown in Table 1.

Table 1 Category Distribution of Network System [根据数据库绘制]

|  |  |  |  |
| --- | --- | --- | --- |
| Invasion Categories | optdigits.tra | optdigits.tes | Total Number of Samples |
|  |  |  |  |
|  |  |  |  |
| Total number of samples in total |  |  | 4143 |

|  |  |
| --- | --- |
| **Abstract**: This data was collected from text ads found on twelve websites that deal with various farm animal related topics. The binary labels are based on whether or not the content owner approves of the ad. |  |

**Source:**

Chris Mesterharm and Michael J. Pazzani   
Rutgers, The State University of New Jersey   
mesterha **'@'** cs.rutgers.edu

**Data Set Information:**

This data was collected from text ads found on twelve websites that deal with various farm animal related topics. Information from the ad creative and the ad landing page is included. The binary labels are based on whether or not the content owner approves of the ad.   
  
For each ad, we include the words on the ad creative and the words from the landing page. Each word from the creative is given a prefix   
of 'ad-'. Title and header HTML markups are noted in a similar way in the text of the landing page. We have already performed stemming and   
stop word removal. Each ad is on a single line. The first word in the line is the label of the instance. It is 1 for accepted ads and -1 for rejected ads.   
  
We have also included a straightforward bag-of-words representation of our data. We use the SVMlight sparse vector format. The first value   
is the label followed by every nonzero attribute. Each of these attributes is encoded as index:value. This is the representation used for the relevant paper cited below.

**Attribute Information:**

Text words in file farm-ads. SVMlight format sparse vectors in file farm-ads-vect.

**Relevant Papers:**

Active Learning using On-line Algorithms. Chris Mesterharm, Michael J. Pazzani. In KDD 2011.