**79.Syskill and Webert Web Page Ratings**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Syskill+and+Webert+Web+Page+Ratings

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 |  |  | 332 |

Michael Pazzani   
Department of Information and Computer Science,   
University of California, Irvine 

**Data Set Information:**

The HTML source of a web page is given. Users looked at each web page and inidated on a 3 point scale (hot medium cold) 50-100 pages per domain. However, this is realistic because we want to learn user profiles from as few examples as possible so that users have an incentitive to rate pages.

**Attribute Information:**

Each subject is in a separate directory. Within each directory, there is an file named "index". The index contains information on the other files. Each entry is a line of the form:   
  
file-name | rating | url | date-rated | title   
  
where file-name is the name of a file (usually an integer), rating is hot, medium, or cold. There are so few medium's that mediums are usually merged with cold in experiments.   
  
The other fields aren't used in learning, but they are collected by the interface for other purposes. They are the url of the html source, the date rated and the title of the web oage.

**Relevant Papers:**

Pazzani M., Billsus, D. (1997). Learning and Revising User Profiles: The identification of interesting web sites. Machine Learning 27, 313-331   
[[Web Link]](http://rexa.info/paper/696ca02710344d1198615399d9117bd49bab8a45)   
  
Pazzani, M., Muramatsu J., Billsus, D. (1996). Syskill & Webert: Identifying interesting web sites. Proceedings of the National Conference on Artificial Intelligence, Portland, OR. PDF

## Data Format

Each subject is in a seperate directory. Within each directory, there is an file named "index". The index contains information on the other files. Each entry is a line of the form:

file-name | rating | url | date-rated | title

where file-name is the name of a file (usually an integer), rating is hot, medium, or cold. There are so few medium's that mediums are usually merged with cold in experiments.

The other fields aren't used in learning, but they are collected by the interface for other purposes. They are the url of the html source, the date rated and the title of the web oage.