**83.IPUMS Census Database**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/IPUMS+Census+Database

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 |  |  | 256932 |

|  |  |
| --- | --- |
| **Abstract**: This data set contains unweighted PUMS census data from the Los Angeles and Long Beach areas for the years 1970, 1980, and 1990. |  |

IPUMS   
Historical Census Projects   
University of Minnesota 

**Data Set Information:**

The original source for this data set is the IPUMS project (RugglesSobek, 1997). The IPUMS project is a large collection of federal census data which has standardized coding schemes to make comparisons across time easy.   
  
The data is an unweighted 1 in 100 sample of responses from the Los Angeles -- Long Beach area for the years 1970, 1980, and 1990. The household and individual records were flattened into a single table and we used all variables that were available for all three years. When there was more than one version of a variable, such as for race, we used the most general. For occupation and industry we used the 1950 basis.   
  
Note that PUMS data is based on cluster samples, i.e. samples are made of households or dwellings from which there may be multiple individuals. Individuals from the same household are no longer independent. Ruggles (1995) considers this issue further and discusses its effect (along with the effects of stratification) on standard errors.   
  
The variable schltype appears to have different coding values across the years 1970, 1980, and 1990.   
  
There are two versions of this data set:   
  
1. The Small Data Set   
  
The small data set contains a 1 in 1000 sample of the Los Angeles and Long Beach area. It was formed by sampling from the large data set.   
  
2. The Large Data Set   
  
The large data set contains a 1 in 100 sample of the Los Angeles and Long Beach area.

**Attribute Information:**

Please see ipums.la.names

This data set contains unweighted PUMS census data from the Los Angeles and Long Beach areas for the years 1970, 1980, and 1990. The coding schemes have been standardized (by the IPUMS project) to be consistent across years.

**The Small Data Set**

The small data set contains a 1 in 1000 sample of the Los Angeles and Long Beach area. It was formed by sampling from the large data set.

**The Large Data Set**

The large data set contains a 1 in 100 sample of the Los Angeles and Long Beach area.