**48.UJI Pen Characters**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/UJI+Pen+Characters

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 D. Llorens, F. Prat, A. Marzal, J. M. Vilar, from departamento de Lenguajes y Sistemas Informaticos, Universitat Jaume I at June 2007.【2.用于[什么实验目的]】We create a character database by collecting samples from 11 writers. Each writer contributed with letters (lower and uppercase), digits, and other characters (Spanish diacritics and punctuation marks) that we have not employed in our experiments and are not included in this database version. The handwriting samples were collected on a Toshiba Portégé M400 Tablet PC using its cordless stylus. 【3】Two samples have been collected for each pair writer/character, so the total number of samples in this database version is 1364, contains 11 writers x 2 repetitions x (2x26 letters + 10 digits). The proposed task is a writer-independent one consisting of 11 leaving-one-writer-out tests, so the effective training set size (for each one of the 1364 test samples) is 1240: 10 writers x 2 repetitions x (2x26 letters + 10 digits). Moreover, this classification task is a 35-class one because we have not considered a different class for each different character: each one of the 26 letters is considered as a case-independent class, there are 9 additional clases for non-zero digits, and the zero is included in the same class as o's. This database is available in a UNIPEN-like format, trying to mimic the original Pendigits database. Two versions of that database are available; see folder: The distribution of our database consists of 12 files:uji.names. One file "UJIpenchars-wNN" per writer, where NN = "01", "02"... "11".【4】The database has 1364 samples, respectively belong to Class[A-N/P-Z] with 44 samples, Class[O] with 66 and Class[0-9] with 22 samples. The categories of system include 36 categories, as shown in Table 1.

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

|  |  |  |
| --- | --- | --- |
| Invasion Categories | Number of samples in each file | Total Number of Samples |
| Class [A] | 4 | 44 |
| Class [B] | 4 | 44 |
| Class [C] | 4 | 44 |
| Class [D] | 4 | 44 |
| Class [E] | 4 | 44 |
| Class [F] | 4 | 44 |
| Class [H] | 4 | 44 |
| Class [I] | 4 | 44 |
| Class [J] | 4 | 44 |
| Class [K] | 4 | 44 |
| Class [L] | 4 | 44 |
| Class [M] | 4 | 44 |
| Class [N] | 4 | 44 |
| Class [O] | 6 | 66 |
| Class [P] | 4 | 44 |
| Class [Q] | 4 | 44 |
| Class [R] | 4 | 44 |
| Class [S] | 4 | 44 |
| Class [T] | 4 | 44 |
| Class [U] | 4 | 44 |
| Class [V] | 4 | 44 |
| Class [W] | 4 | 44 |
| Class [X] | 4 | 44 |
| Class [Y] | 4 | 44 |
| Class [Z] | 4 | 44 |
| Class [1] | 2 | 22 |
| Class [2] | 2 | 22 |
| Class [3] | 2 | 22 |
| Class [4] | 2 | 22 |
| Class [5] | 2 | 22 |
| Class [6] | 2 | 22 |
| Class [7] | 2 | 22 |
| Class [8] | 2 | 22 |
| Class [9] | 2 | 22 |
| Total number of samples in total | 124 | 1364 |

3. 精简描述

The Data in our experiment were collected by D. Llorens, F. Prat, A. Marzal, J. M. Vilar, from departamento de Lenguajes y Sistemas Informaticos, Universitat Jaume I, at June, 2007. The dataset includes have 1364 samples, which used to classify the UJI Pen characters. Through which, we divided the dataset into two part, training data set with 1000 samples and forecasting data set with 364 samples.