**47.Pen-Based Recognition of Handwritten Digits**

1. 数据库网址

https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits

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, Fevzi. Alimoglu, from department of Computer Engineering, Bogazici University at July 1998.【2.用于[什么实验目的]】We create a digit database by collecting 250 samples from 44 writers. The samples written by 30 writers are used for training, cross-validation and writer dependent testing, and the digits written by the other 14 are used for writer independent testing. This database is also available in the UNIPEN format.【3】These writers are asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution. Subject are monitored only during the first entry screens. Each screen contains five boxes with the digits to be written displayed above. Subjects are told to write only inside these boxes. If they make a mistake or are unhappy with their writing, they are instructed to clear the content of a box by using an on-screen button. The first ten digits are ignored because most writers are not familiar with this type of input devices, but subjects are not aware of this. Number of Attributes is 16 input+1 class attribute, all input attributes are integers in the range 0-100. The last attribute is the class code 0-9. In our study, we use only ($x, y$) coordinate information. The stylus pressure level values are ignored. First we apply normalization to make our representation invariant to translations and scale distortions. The raw data that we capture from the tablet consist of integer values between 0 and 500 (tablet input box resolution). The new coordinates are such that the coordinate which has the maximum range varies between 0 and 100. Usually $x$ stays in this range, since most characters are taller than they are wide.【4】The database has 10992 samples, respectively belong to optdigits.tra with 3823 samples and optidigits.tes with 1797 samples. The way we used the dataset was to use first half of training for actual training, one-fourth for validation and one-fourth for writer-dependent testing. The test set was used for writer-independent testing and is the actual quality measure. The categories of system include two categories, as shown in Table 1.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Categories | pendigits.tra | pendigits.tes | Total Number of Samples |
| 0 | 780 | 363 | 1143 |
| 1 | 779 | 364 | 1143 |
| 2 | 780 | 364 | 1144 |
| 3 | 719 | 336 | 1054 |
| 4 | 780 | 364 | 1144 |
| 5 | 720 | 335 | 1055 |
| 6 | 720 | 336 | 1056 |
| 7 | 778 | 364 | 1144 |
| 8 | 719 | 336 | 1054 |
| 9 | 719 | 336 | 1055 |
| Total number of samples in total | 7494 | 3498 | 10992 |

3. 精简描述

The Data in our experiment were collected by E. Alpaydin, Fevzi. Alimoglu, from department of Computer Engineering, Bogazici University at July,1998. The dataset includes have 5620 samples, which used to for training, cross-validation and writer dependent testing, and the digits written by the other 14 are used for writer independent testing. This database is also available in the UNIPEN format. Through which, we divided the dataset into two part, training data set with 7494 samples and forecasting data set with 3498 samples.