**71.ISOLET**

1. 数据库网址

https://archive.ics.uci.edu/ml/datasets/ISOLET

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 Ron Cole and Mark Fanty, from department of Computer Science and Engineering, Oregon Graduate Institute.【2.用于[什么实验目的]】We used to classify the letters from the word.【3】This data set was generated as follows. 150 subjects spoke the name of each letter of the alphabet twice. Hence, we have 52 training examples from each speaker. The speakers are grouped into sets of 30 speakers each, and are referred to as isolet1, isolet2, isolet3, isolet4, and isolet5. The data appears in isolet1+2+3+4.data in sequential order, first the speakers from isolet1, then isolet2, and so on. The test set, isolet5, is a separate file. The features are described in the paper by Cole and Fanty cited above. The features include spectral coefficients; contour features, sonorant features, pre-sonorant features, and post-sonorant features. Exact order of appearance of the features is not known. The features are described in the paper by Cole and Fanty cited above. The features include spectral coefficients; contour features, sonorant features, pre-sonorant features, and post-sonorant features. Exact order of appearance of the features is not known. All attributes are continuous, real-valued attributes scaled into the range -1.0 to 1.0. 【4】The database has 5620 samples, respectively belong to isolet1+2+3+4.data.Z with 6238 samples and isolet5.data.Z with 1559 samples. As shown in Table 1.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Categories | Isolet1+2+3+4 | Isolet5 | Total Number of Samples |
| 1 A | 240 | 60 | 300 |
| 2 B | 240 | 60 | 300 |
| 3 C | 240 | 60 | 300 |
| 4 D | 240 | 60 | 300 |
| 5 E | 240 | 60 | 300 |
| 6 F | 238 | 60 | 298 |
| 7 G | 240 | 60 | 300 |
| 8 H | 240 | 60 | 300 |
| 9 I | 240 | 60 | 300 |
| 10 J | 240 | 60 | 300 |
| 11 K | 240 | 60 | 300 |
| 12 L | 240 | 60 | 300 |
| 13 M | 240 | 59 | 299 |
| 14 N | 240 | 60 | 300 |
| 15 O | 240 | 60 | 300 |
| 16 P | 240 | 60 | 300 |
| 17 Q | 240 | 60 | 300 |
| 18 R | 240 | 60 | 300 |
| 19 S | 240 | 60 | 300 |
| 20 T | 240 | 60 | 300 |
| 21 U | 240 | 60 | 300 |
| 22 V | 240 | 60 | 300 |
| 23 W | 240 | 60 | 300 |
| 24 X | 240 | 60 | 300 |
| 25 Y | 240 | 60 | 300 |
| 26 Z | 240 | 60 | 300 |
| Total number of samples in total | 6238 | 1559 | 7797 |

3. 精简描述

The Data in our experiment were collected by Ron Cole and Mark Fanty, from department of Computer Science and Engineering, Oregon Graduate Institute. The dataset includes have 7797 samples, which used to classify the letters from the word. Through which, we divided the dataset into two part, training data set with 6238 samples and forecasting data set with 1559 samples.