**62.Dexter**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Dexter

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 Thorsten Joachims in the “bag-of-words” representation.【2.用于[什么实验目的]】We used to classify the word in the text.【3】There were 9947 features (of which 2562 are always zeros for all the examples) representing frequencies of occurrence of word stems in text. The task is to learn which Reuters articles are about 'corporate acquisitions'. We added a number of distractor feature called 'probes' having no predictive power. The order of the features and patterns were randomized. Number of variables/features/attributes: Real is 9947, Probes is 10053, and Total is 20000. The data are in the following format is dataname.param, dataname.feat, dataname\_train.data, dataname\_valid.data, dataname\_test.data, dataname\_train.labels, dataname\_valid.labels, dataname\_test.labels.【4】The database has 2600 samples, respectively belong to positive with 1300 samples and negative with 1300 samples. The categories can be as shown in Table 1.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Invasion Categories | Positive ex. | Negative ex. | Total Number of Samples |
| Training | 150 | 150 | 300 |
| Validation | 150 | 150 | 300 |
| Test | 1000 | 1000 | 2000 |
| Total number of samples in total | 1300 | 1300 | 2600 |

3. 精简描述

The Data in our experiment were collected by Thorsten Joachims in the “bag-of-words” representation. The dataset includes have 2600 samples, which used to classify the word in the text. Through which, we divided the dataset into two part, training data set with 1300 samples and forecasting data set with 1300 samples.