**96.CNAE-9 – 巴西公司分类**

1. 数据库网址

https://archive.ics.uci.edu/ml/datasets/CNAE-9

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 |  |  | 1080 |

|  |  |
| --- | --- |
| **Abstract**: This is a data set containing 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories |  |

**Source:**

Patrick Marques Ciarelli, pciarelli **'@'** lcad.inf.ufes.br, Department of Electrical Engineering, Federal University of Espirito Santo   
Elias Oliveira, elias **'@'** lcad.inf.ufes.br, Department of Information Science, Federal University of Espirito Santo

**Data Set Information:**

This is a data set containing 1080 documents of free text business descriptions of Brazilian companies categorized into a   
subset of 9 categories cataloged in a table called National Classification of Economic Activities (ClassificaÃ§Ã£o Nacional de   
Atividade EconÃ´micas - CNAE). The original texts were pre-processed to obtain the current data set: initially, it was kept only   
letters and then it was removed prepositions of the texts. Next, the words were transformed to their canonical form. Finally,   
each document was represented as a vector, where the weight of each word is its frequency in the document. This data set is   
highly sparse (99.22% of the matrix is filled with zeros).

**Attribute Information:**

In the data set there are 857 attributes, 1 attributes with the class of instance and 856 with word frequency:   
1. category: range 1 - 9 (integer)   
2 - 857. word frequency: (integer)

3. Past Usage:

- Patrick Marques Ciarelli, Elias Oliveira, "Agglomeration and Elimination of Terms for Dimensionality Reduction",

Ninth International Conference on Intelligent Systems Design and Applications, pp. 547-552, 2009:

- Feature selection (900 instances for training and 180 instances for test):

- Best results using kNN (k=1):

50 dimensions: 87.78% (LSI)

100 dimensions: 92.78% (LSI)

150 dimensions: 92.22% (LSI)

200 dimensions: 92.78% (MI)

250 dimensions: 92.78% (MI)

- Patrick Marques Ciarelli, Elias Oliveira, Evandro O. T. Salles, "An Evolving System Based on Probabilistic Neural Network",

Brazilian Symposium on Artificial Neural Network, 2010:

- Incremental learning (no off-line training):

- Best result:

88.71% (ePNN)

4. Relevant Information:

- This is a data set containing 1080 documents of free text business descriptions of Brazilian companies categorized into a

subset of 9 categories cataloged in a table called National Classification of Economic Activities (Classifica��o Nacional de

Atividade Econ�micas - CNAE). The original texts were pre-processed to obtain the current data set: initially, it was kept only

letters and then it was removed prepositions of the texts. Next, the words were transformed to their canonical form. Finally,

each document was represented as a vector, where the weight of each word is its frequency in the document. This data set is

highly sparse (99.22% of the matrix is filled with zeros).

5. Number of Instances: 1080

6. Number of Attributes: 857 (1 category, 856 word frequency)

7. Attribute Information:

1. category: range 1 - 9 (integer)

2 - 857. word frequency: (integer)

8. Missing Attribute Values: None

9. Class Distribution: the categories are equally distribuited. (120 instances in each of nine categories)

Summary Statistics:

Min Max Mean SD

word frequency: 0 4 0.0082 0.0948