

Benchmark data for classification

1. Machine Learning Domain

- UCI Machine Learning Repository
<http://www.ics.uci.edu/mlearn/MLRepository.html>
Most important machine learning benchmark datasets. Including many dataset for classification. For example, Boston Housing data.
Some of the datasets are also available in R's package: mlbench: Machine Learning Benchmark Problems
<http://cran.r-project.org>
Search for package: mlbench.
Possible problem with the UCI Machine Learning Repository is that all the datasets are hard to overfit using tree-structured methods. (Mark R. Segal, 2004) This raises issues about the scope of the repository.
- UCI Knowledge Discovery in Databases Archive
<http://kdd.ics.uci.edu/>
- Handwritten Recognition: Digits, alphabets, address recognition,...

2. DNA Microarray Domain

Leukemia, Colon cancer, Lymphoma

3. Statlog Project <http://www.liacc.up.pt/ML/statlog/>

This project was concerned with comparative studies of different machine learning, neural and statistical classification algorithms. About 20 different algorithms were evaluated on more than 20 different datasets.

4. General Machine Learning Resource Website <http://www.cs.ust.hk/ivor/resource.htm>

Many machine learning source codes available, as well as many datasets.