

UCI Databases Statistics

Mauricio Kugler

October 18, 2004

Contents

1	Introduction	1
2	PRD File Format	1
3	Databases Modifications	2
4	Databases Statistics	2

1 Introduction

This document presents the statistics of some of the UCI databases [BM98]: *bupa*, *dermatology*, *ecoli*, *glass*, *ionosphere*, *iris*, *isolet*, *letter*, *lrs*, *lung*, *pendigits*, *pima*, *satimage*, *segment*, *sonar*, *vehicle*, *vowel* and *wine*. That statistics are based on the files available in the authors web page¹. The files were converted to the PRD data format described in section 2. Also, some databases were slightly modified, and the modification and reasons for them are explained in section 3. Finally, the databases statistics are presented in section 4.

2 PRD File Format

The PRD (Pattern Recognition Data) is an ASCII file format created to make the databases easy to read in C/C++ programs. It is defined as follows (where each line begins with the line number just for reference):

```
1      PRD
2      <number of samples>
3      <number of features>
4      s [0] [0] , s [0] [1] , ... , s [0] [n-2] , s [0] [n-1] , <label>
5      s [1] [0] , s [1] [1] , ... , s [1] [n-2] , s [1] [n-1] , <label>
6      s [2] [0] , s [2] [1] , ... , s [2] [n-2] , s [2] [n-1] , <label>
...
m+2    s [m-2] [0] , s [m-2] [1] , ... , s [m-2] [n-2] , s [m-2] [n-1] , <label>
m+3    s [m-1] [0] , s [m-1] [1] , ... , s [m-1] [n-2] , s [m-1] [n-1] , <label>
```

¹<http://www.mauricio.kugler.com/database.html>

where m is the number of samples, n is the number of features and $s[a][b]$ means the feature b of sample a .

The class label must be an integer and the sequences of classes labels cannot be disrupted. For example, in a 10 classes database, the classes must be labeled from 0 to 9, but they can appear on the file at any order. The samples features are divided by commas with no spaces before or after it. The lines are numbered here just for reference, but they are not numbered in the databases files. Finally, there are no spaces after the three header lines.

3 Databases Modifications

To eliminate constant features and missing values samples and/or features, the following modifications were made in the UCI databases:

- For the *dermatology* database, the 8 patterns with missing values were removed from the database, changing the samples number from 366 to 358;
- The *glass* database have one of the 7 classes with no patterns. This class was not considered in the PRD file class labeling. So, the labels are from 0 to 5;
- In the *ionosphere* database, the second feature of is constant equal to 0 and had been removed, changing the number of features from 34 to 33;
- The letter database had been split in training and test data with the first 12200 samples for training and the remaining 7800 for test;
- The classes in the original lrs database are numbered from 0 to 99, but this numbering are not continuous. The 48 present classes were relabeled from 0 to 47. Also, the 10 "header" features were eliminated, changing the feature number from 103 to 93;
- For the lung database, the 26th sample and the 5th feature were removed because missing values, changing the samples number from 32 to 31 and the features number from 56 to 55;
- The satimage database have one of the 7 classes with no patterns. This class was not considered in the PRD file class labeling. So, the labels are from 0 to 5;
- For the segment database, the feature that contains the number of pixels ("region-pixel-count") is constant (always 9) and was removed, changing the original number of features from 19 to 18;

4 Databases Statistics

The summary of the databases are given in Table 1. After, the complete statistics of class distribution of all databases are given.

Table 1: Databases summary

Problem	Trn. Data	Tst. Data	Classes	Features
<i>bupa</i>	345	-	2	6
<i>dermatology</i>	358	-	6	34
<i>ecoli</i>	336	-	8	7
<i>glass</i>	214	-	6	9
<i>ionosphere</i>				
<i>iris</i>	150	-	3	4
<i>isolet</i>	6238	1559	26	617
<i>letter</i>	12200	7800	26	16
<i>lrs</i>	531	-	48	93
<i>lung</i>	31	-	3	55
<i>pendigits</i>	7494	3498	10	16
<i>pima</i>	768	-	2	8
<i>satimage</i>	4435	2000	6	36
<i>segment</i>	210	2100	7	18
<i>sonar</i>	208	-	2	60
<i>vehicle</i>	846	-	4	18
<i>vowel</i>	528	462	11	11
<i>wine</i>	178	-	3	13

bupa(Original name: *BUPA liver disorders*)

Class	Samples	Distributions	Original Name
0	145	42.03%	1
1	200	57.97%	2

dermatology(Original name: *Dermatology Database*)

Class	Samples	Distributions	Original Name
0	111	31.01%	1
1	60	16.76%	2
2	71	19.83%	3
3	48	13.41%	4
4	48	13.41%	5
5	20	5.59%	6

ecoli

(Original name: *Protein Localization Sites*)

Class	Samples	Distributions	Original Name
0	143	42.56%	cp
1	77	22.92%	im
2	52	15.48%	pp
3	35	10.42%	imU
4	20	5.95%	om
5	5	1.49%	omL
6	2	0.60%	imL
7	2	0.60%	imS

glass

(Original name: *Glass Identification Database*)

Class	Samples	Distributions	Original Name
0	70	32.71%	1
1	76	35.51%	2
2	17	7.94%	3
3	13	6.07%	5
4	9	4.21%	6
5	29	13.55%	7

ionosphere

(Original name: *Johns Hopkins University Ionosphere database*)

Class	Samples	Distributions	Original Name
0	225	64.10%	g
1	126	35.90%	b

iris

(Original name: *Iris Plant Database*)

Class	Samples	Distributions	Original Name
0	50	33.33%	Iris Setosa
1	50	33.33%	Iris Versicolour
2	50	33.33%	Iris Virginica

isolet

(Original name: *Isolated Letter Speech Recognition*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	240	3.79%	60	3.85%	A
1	240	3.79%	60	3.85%	B
2	240	3.79%	60	3.85%	C
3	240	3.79%	60	3.85%	D
4	240	3.79%	60	3.85%	E
5	238	3.76%	60	3.85%	F
6	240	3.79%	60	3.85%	G
7	240	3.79%	60	3.85%	H
8	240	3.79%	60	3.85%	I

9	240	3.79%	60	3.85%	J
10	240	3.79%	60	3.85%	K
11	240	3.79%	60	3.85%	L
12	240	3.79%	59	3.78%	M
13	240	3.79%	60	3.85%	N
14	240	3.79%	60	3.85%	O
15	240	3.79%	60	3.85%	P
16	240	3.79%	60	3.85%	Q
17	240	3.79%	60	3.85%	R
18	240	3.79%	60	3.85%	S
19	240	3.79%	60	3.85%	T
20	240	3.79%	60	3.85%	U
21	240	3.79%	60	3.85%	V
22	240	3.79%	60	3.85%	W
23	240	3.79%	60	3.85%	X
24	240	3.79%	60	3.85%	Y
25	240	3.79%	60	3.85%	Z

letter

(Original name: *Letter Image Recognition Data*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	489	4.01%	300	3.85%	A
1	466	3.82%	300	3.85%	B
2	436	3.57%	300	3.85%	C
3	505	4.14%	300	3.85%	D
4	468	3.84%	300	3.85%	E
5	475	3.89%	300	3.85%	F
6	473	3.88%	300	3.85%	G
7	434	3.56%	300	3.85%	H
8	455	3.73%	300	3.85%	I
9	447	3.66%	300	3.85%	J
10	439	3.60%	300	3.85%	K
11	461	3.78%	300	3.85%	L
12	492	4.03%	300	3.85%	M
13	483	3.96%	300	3.85%	N
14	453	3.71%	300	3.85%	O
15	503	4.12%	300	3.85%	P
16	483	3.96%	300	3.85%	Q
17	458	3.75%	300	3.85%	R
18	448	3.67%	300	3.85%	S
19	496	4.07%	300	3.85%	T
20	513	4.20%	300	3.85%	U
21	464	3.80%	300	3.85%	V
22	452	3.70%	300	3.85%	W
23	487	3.99%	300	3.85%	X
24	486	3.98%	300	3.85%	Y

25 434 3.56% 300 3.85% Z

lrs

(Original name: *Part of the IRAS
Low Resolution Spectrometer Database*)

Class	Samples	Distributions	Original Name
0	3	0.56%	2
1	1	0.19%	3
2	7	1.32%	4
3	1	0.19%	5
4	2	0.38%	12
5	3	0.56%	13
6	10	1.88%	14
7	24	4.52%	15
8	13	2.45%	16
9	12	2.26%	17
10	26	4.90%	18
11	15	2.82%	21
12	27	5.08%	22
13	31	5.84%	23
14	19	3.58%	24
15	19	3.58%	25
16	30	5.65%	26
17	35	6.59%	27
18	42	7.91%	28
19	55	10.36%	29
20	9	1.69%	31
21	10	1.88%	32
22	4	0.75%	33
23	4	0.75%	34
24	3	0.56%	35
25	1	0.19%	36
26	1	0.19%	37
27	3	0.56%	38
28	3	0.56%	39
29	8	1.51%	41
30	45	8.47%	42
31	20	3.77%	43
32	18	3.39%	44
33	5	0.94%	45
34	1	0.19%	50
35	2	0.38%	69
36	1	0.19%	71
37	1	0.19%	72

38	1	0.19%	73
39	3	0.56%	79
40	3	0.56%	80
41	2	0.38%	81
42	1	0.19%	82
43	1	0.19%	85
44	3	0.56%	91
45	1	0.19%	92
46	1	0.19%	95
47	1	0.19%	96

lung

(Original name: *Lung Cancer Data*)

Class	Samples	Distributions	Original Name
0	9	29.03%	1
1	13	41.94%	2
2	9	29.03%	3

pendigits

(Original name: *Pen-Based Recognition of Handwritten Digits*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	780	10.41%	363	10.38%	0
1	779	10.39%	364	10.41%	1
2	780	10.41%	364	10.41%	2
3	719	9.59%	336	9.61%	3
4	780	10.41%	364	10.41%	4
5	720	9.61%	335	9.58%	5
6	720	9.61%	336	9.61%	6
7	778	10.38%	364	10.41%	7
8	719	9.59%	336	9.61%	8
9	719	9.59%	336	9.61%	9

pima

(Original name: *Pima Indians Diabetes Database*)

Class	Samples	Distributions	Original Name
0	500	65.10%	0
1	268	34.90%	1

satimage

(Original name: *Landsat Multi-Spectral Scanner image data*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	1072	24.17%	461	23.05%	1
1	479	10.80%	224	11.20%	2
2	961	21.67%	397	19.85%	3
3	415	9.36%	211	10.55%	4
4	470	10.60%	237	11.85%	5
5	1038	23.40%	470	23.50%	7

segment

(Original name: *Image Segmentation data*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	30	14.29%	301	14.33%	Brickface
1	30	14.29%	300	14.29%	Sky
2	30	14.29%	296	14.10%	Foliage
3	30	14.29%	300	14.29%	Cement
4	30	14.29%	299	14.24%	Window
5	30	14.29%	300	14.29%	Path
6	30	14.29%	297	14.14%	Grass

sonar

(Original name: *Sonar, Mines vs. Rocks*)

Class	Samples	Distributions	Original Name
0	97	46.63%	R
1	111	53.37%	M

vowel

(Original name: *Vowel Recognition - Deterding data*)

Class	Trn. Samples	Trn. Dtrb.	Tst. Samples	Tst. Dtrb.	Original Name
0	48	9.09%	42	9.09%	0
1	48	9.09%	42	9.09%	1
2	48	9.09%	42	9.09%	2
3	48	9.09%	42	9.09%	3
4	48	9.09%	42	9.09%	4
5	48	9.09%	42	9.09%	5
6	48	9.09%	42	9.09%	6
7	48	9.09%	42	9.09%	7
8	48	9.09%	42	9.09%	8
9	48	9.09%	42	9.09%	9
10	48	9.09%	42	9.09%	10

vehicle

(Original name: *Vehicle Silhouettes*)

Class	Samples	Distributions	Original Name
0	199	23.52%	van
1	217	25.65%	saab
2	218	25.77%	bus
3	212	25.06%	opel

wine

(Original name: *Wine Recognition Database*)

Class	Samples	Distributions	Original Name
0	59	33.15%	1
1	71	39.89%	2
2	48	26.97%	3

References

- [BM98] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.