1. Title of Database: ISOLET (Isolated Letter Speech Recognition)
2. Sources:
   (a) Creators: Ron Cole and Mark Fanty
       Department of Computer Science and Engineering,
       Oregon Graduate Institute, Beaverton, OR 97006.
       cole@cse.ogi.edu, fanty@cse.ogi.edu
   (b) Donor: Tom Dietterich
       Department of Computer Science
       Oregon State University, Corvallis, OR 97331
       tgd@cs.orst.edu
   (c) September 12, 1994

3. Past Usage:
   (a) Fanty, M., Cole, R. (1991). Spoken letter recognition. In
       Lippman, R. P., Moody, J., and Touretzky, D. S. (Eds).
       Advances in Neural Information Processing Systems 3. San
       Mateo, CA: Morgan Kaufmann.

       Goal: Predict which letter-name was spoken--a simple
       classification task. 95.9% correct classification using
       the OPT backpropagation implementation. Training on
       isolet1+2+3+4, testing on isolet5. Network architecture: 56
       hidden units, 26 output units (one-per-class).

   (b) Dietterich, T. G., Bakiri, G. (1991) Error-correcting
       output codes: A general method for improving multiclass
       inductive learning programs. Proceedings of the Ninth
       National Conference on Artificial Intelligence (AAAI-91),
       Anaheim, CA: AAAI Press.

       Goal: same as above. 95.83% correct using OPT
       backpropagation. (Architecture: 78 hidden units, 26 output
       units, one-per-class).
       96.73% correct using a 30-bit error-correcting output code with
       OPT (Architecture: 156 hidden units, 30 output units).

   (c) Dietterich, T. G., Bakiri, G. (1994) Solving Multiclass
       Learning Problems via Error-Correcting Output Codes.
       Submitted. Available as URL
       ftp://ftp.cs.orst.edu/pub/tgd/papers/tr-ecoc.ps.gz

       Supporting data not published in that paper:

Algorithm and configuration	errors	%error	%correct
Opt 30-bit ECOC	51	3.27	96.73
Opt 62-bit ECOC	63	4.04	95.96
Opt OPC	65	4.17	95.83
C4.5 107-bit ECOC soft pruned	103	6.61	93.39
C4.5 92-bit ECOC soft pruned	107	6.86	93.14
C4.5 45-bit ECOC soft pruned	109	6.99	93.01
C4.5 107-bit ECOC soft raw	116	7.44	92.56
C4.5 92-bit ECOC soft raw	118	7.57	92.43
C4.5 107-bit ECOC hard pruned	126	8.08	91.91
C4.5 92-bit ECOC hard pruned	127	8.15	91.85
C4.5 62-bit ECOC soft pruned	131	8.40	91.60
C4.5 30-bit ECOC soft pruned	134	8.60	91.40
C4.5 62-bit ECOC soft raw	134	8.60	91.40
C4.5 77-bit ECOC hard pruned	138	8.85	91.15

C4.5 45-bit ECOC soft raw	145	9.30	90.70
C4.5 62-bit ECOC hard pruned	164	9.88	90.12
C4.5 45-bit ECOC hard pruned	155	9.94	90.06
C4.5 30-bit ECOC soft raw	175	11.23	88.77
C4.5 30-bit ECOC hard pruned	185	11.87	88.13
C4.5 multiclass soft pruned	239	15.33	84.67
C4.5 multiclass soft raw	248	15.91	84.09
C4.5 multiclass hard pruned	254	16.29	83.71
C4.5 15-bit ECOC soft pruned	259	16.61	83.39
C4.5 multiclass hard raw	264	16.93	83.07
C4.5 OPC soft pruned	296	18.99	81.01
C4.5 15-bit ECOC soft raw	321	20.59	79.41
C4.5 107-bit ECOC hard raw	334	21.42	78.58
C4.5 92-bit ECOC hard raw	349	22.39	77.61
C4.5 OPC soft raw	379	24.31	75.69
C4.5 15-bit ECOC hard pruned	383	24.57	75.43
C4.5 77-bit ECOC hard raw	424	27.20	72.80
C4.5 OPC hard pruned	437	28.03	71.97
C4.5 62-bit ECOC hard raw	463	29.70	70.30
C4.5 OPC hard raw	519	33.29	66.71
C4.5 45-bit ECOC hard raw	568	36.43	63.57
C4.5 30-bit ECOC hard raw	617	43.04	56.96
C4.5 15-bit ECOC hard raw	991	63.57	36.43

Legend: OPT = conjugate-gradient implementation of backprop.
C4.5 = Quinlan's C4.5 system, Release 1.
OPC = one-per-class representation
ECOC = error-correcting output code
raw = unpruned decision trees
pruned = pruned decision trees (CF=0.25)
hard = default trees
soft = trees with softened thresholds.
multiclass = one tree to do all 26-way classifications.

4. Relevant Information Paragraph:

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.

You will note that 3 examples are missing. I believe they were
dropped due to difficulties in recording.

I believe this is a good domain for a noisy, perceptual task. It
is also a very good domain for testing the scaling abilities of
algorithms. For example, C4.5 on this domain is slower than
backpropagation!

I have formatted the data for C4.5 and provided a C4.5-style
names file as well.

5. Number of Instances

isolet1+2+3+4.data.Z: 6238
isolet5.data.Z: 1559

6. Number of Attributes 617 plus 1 for the class
   All attributes are continuous, real-valued attributes scaled into the range -1.0 to 1.0.

7. For Each Attribute: (please give both acronym and full name if both exist)
   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.

8. Missing Attribute Values: none

9. Class Distribution:

Class		isolet1+2+3+4:	isolet5:
1	A	240	60
2	B	240	60
3	C	240	60
4	D	240	60
5	E	240	60
6	F	238	60
7	G	240	60
8	H	240	60
9	I	240	60
10	J	240	60
11	K	240	60
12	L	240	60
13	M	240	59
14	N	240	60
15	O	240	60
16	P	240	60
17	Q	240	60
18	R	240	60
19	S	240	60
20	T	240	60
21	U	240	60
22	V	240	60
23	W	240	60
24	X	240	60
25	Y	240	60
26	Z	240	60