Comparison of the Cover Tree and sb(S) Datastructures Additional Experiments

January 3, 2005

	Cover Tree		sb(S)	
Dataset	Total	Construction	Total	Construction
iris	0.000959	0.000604	0.001187	0.000390
bupa	0.003925	0.001533	0.005595	0.001175
wine	0.001209	0.000792	0.001595	0.000473
glass	0.001657	0.001002	0.002645	0.000552
pima	0.008156	0.003601	0.015552	0.004002
ionosphere	0.005082	0.002578	0.022682	0.003272
pendigits.tes	0.068224	0.022312	0.163722	0.059033
optdigits.tes	0.171117	0.034527	0.275410	0.170200
pendigits.tra	0.233144	0.056984	0.583384	0.235772
optdigits.tra	0.684487	0.098905	1.141936	0.806184
letter	2.671169	0.198672	5.623558	3.423026
corel	20.811123	0.541603	28.764572	17.670626
phy_train	7.885545	1.180012	159.033859	1.521987
bio_test	461.434052	8.222633	385.882141	149.366470
bio_train	504.397034	7.421129	374.645599	153.598984
covtype	42.961823	7.681526	237.976212	113.729851
phy_test	21.867506	2.581782	655.451843	4.343706
mnist	1629.700684	256.675537	3002.393066	2715.849854

Table 1: CPU Time (in seconds) spent on preprocessing and 1-nearest neighbor search (effectively self-search) of every point in the dataset (total time).

Dataset	Cover Tree	sb(S)
iris.data	0.001253	0.005672
bupa.data	0.005338	0.039804
wine.data	0.001489	0.006833
glass.data	0.002148	0.017301
pima.data	0.011447	0.110684
ionosphere.data	0.009324	0.124497
pendigits.tes.data	0.104286	1.336359
optdigits.tes.data	0.277466	3.538248
pendigits.tra.data	0.363333	5.291792
optdigits.tra.data	1.185085	14.349384
letter.data	4.833340	78.584114
corel.data	34.991699	426.650787
phy_train.data	14.213729	896.944641
bio_test.data	726.925720	7,688.179199
bio_train.data	788.968384	7,862.303711
covtype.data	69.011177	1,607.416138
phy_test.data	39.006523	3,866.373779

Table 2: Total CPU Time (in seconds) spent on preprocessing and 2-nearest neighbor search of every point.

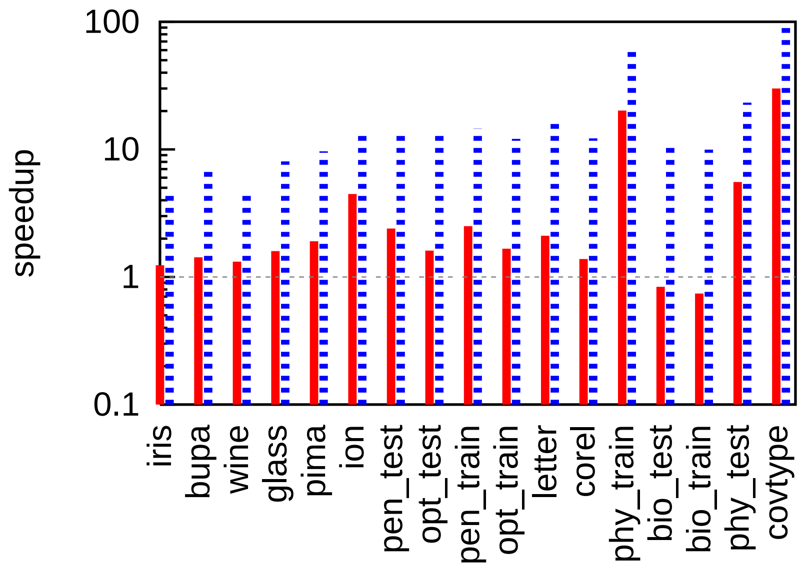

Figure 1: Speedup of the Cover Tree over sb(S) for 1-nearest neighbor search (solid impulses) and 2-nearest neighbor search (dotted impulses) of every point in the dataset. Construction time is included.

		Cover Tree		sb(S)	
Dataset	Size	Construction	Total	Construction	Total
male-names.data	8.0K	0.039979	0.152851	0.137473	0.285567
counties.data	13K	0.093986	0.496662	0.509397	0.924691
verbs.data	21K	0.123313	0.889539	0.878854	1.635793
adj.exc	22K	0.146470	0.890195	0.890796	1.893008
female-names.data	29K	0.201007	1.043289	1.039647	2.195311
verb.exc	37K	0.385958	1.913947	2.938263	6.301516
noun.exc	37K	0.758334	3.208628	4.688563	8.719802
index.adv	51K	0.708111	4.221557	5.776797	12.905276
places.data	96K	1.825724	13.419159	16.682199	34.558277
mcds.data	96K	1.583470	13.853719	21.225595	42.396118
index.adj	213K	5.508421	54.142963	74.936234	150.300598
words.data	433K	10.698040	181.889008	173.332932	351.881256
surnames.data	679K	17.708891	366.990936	337.255219	545.682800
county-cousub2k.data	750K	90.586899	795.614807	1,096.797485	1,391.571411
usenet-names	884K	30.331722	750.781921	1,024.264526	1,764.839722
index.noun	1.4M	273.822510	2,368.036865	4,464.267578	11,221.715820

Table 3: String datasets under edit distance. CPU Time (in seconds) spent on preprocessing and 1-nearest neighbor search of every point.

	Cover Tree		sb(S)	
Dataset	Construction	Total	Construction	Total
iris	1,438	3,777	845	3,650
bupa	5,436	23,051	3,596	19,056
wine	1,670	4,232	819	3,887
glass	2,854	7,184	1,302	8,796
pima	11,130	48,966	12,089	43,080
ionosphere	12,184	28,828	6,136	44,959
pendigits.tes	90,244	441,599	152,096	344,870
optdigits.tes	152,268	813,177	201,639	309,714
pendigits.tra	241,790	1,612,658	534,668	1,134,608
optdigits.tra	440,646	3,035,235	764,478	1,075,958
letter	1,033,794	18,464,245	4,988,272	8,197,917
corel	2,496,568	73,654,332	15,859,193	27,486,178
phy_train	2,954,646	21,047,508	1,087,590	119,064,970
bio_test	25,143,600	892,731,798	90,665,396	237,219,310
bio_train	23,333,460	967,708,624	94,099,866	225,907,992
covtype	16,954,564	130,064,613	83,403,428	187,004,255
phy_test	6,116,334	52,815,495	3,000,138	465,077,203
mnist	128,892,460	796,532,454	304,359,444	336,799,222

Table 4: Number of distance function calls for data structure construction and 1-Nearest Neighbor Search of every point in the dataset.

	Cover Tree		sb(S)	
Dataset	Construction	Total	Construction	Total
iris	1,438	5,496	845	19,894
bupa	5,436	35,234	3,596	128,830
wine	1,670	6,260	819	17,929
glass	2,854	10,415	1,302	58,033
pima	11,130	75,834	12,089	285,780
ionosphere	12,184	56,023	6,136	230,616
pendigits.tes	90,244	699,002	152,096	2,188,879
optdigits.tes	152,268	1,311,960	201,639	3,292,214
pendigits.tra	241,790	2,543,765	534,668	7,556,070
optdigits.tra	440,646	4,866,958	764,478	11,128,762
letter	1,033,794	28,716,320	4,988,272	63,350,102
corel	2,496,568	112,404,509	15,859,193	233,569,010
phy_train	2,954,646	34,850,249	1,087,590	637,542,082
bio_test	25,143,592	1,373,615,602	90,665,396	3,934,506,804
bio_train	23,333,454	1,478,747,580	94,099,866	3,983,924,071
covtype	16,954,564	192,299,944	83,403,428	960,331,925
phy_test	6,116,306	87,449,516	3,000,138	2,534,436,586
mnist	128,892,278	1,632,068,482	304,359,444	1,287,163,309

Table 5: Number of distance function calls for data structure construction and 2-nearest neighbor search of every point in the dataset.

Dataset	Cover Tree	sb(S)	Ratio	Dataset Size
letter	9,744	8,824	1.10	696K
corel	21,840	20,140	1.08	11M
phy_train	45,612	36,544	1.25	49M
bio_test	122,988	109,564	1.12	63M
bio_train	127,916	116,092	1.10	65M
covtype	408,280	384,140	1.06	72M
phy_test	90,376	79,584	1.14	98M
mnist	391,464	382,148	1.02	105M

Table 6: Maximum amount of memory (in Kbytes) claimed by the data structures during construction and 1-Nearest Neighbor Search of every point in the dataset. (The maximum is representative – there are no peaks in memory usage.) Memory was profiled every second using the ps utility. Datasets taking less than 1 sec are not reported. The amount of space required by both datastructures seems comparable, with the cover tree taking at most 25% more space.