Complexity Curve: a Graphical Measure of Data Complexity and Classifier Performance

Supplementary document S1: Data Sets Properties

Julian Zubek[1,2], Dariusz Plewczynski[2]

1. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
2. Centre of New Technologies, University of Warsaw, Warsaw, Poland

Data set	Instances	Attributes	Classes	Imbalance ratio
HDDT BREAST-Y	286	9	2	2.36
HDDT COMPUSTAT	13657	20	2	25.26
HDDT COVTYPE	38500	10	2	13.02
HDDT CREDIT-G	1000	20	2	2.33
HDDT ESTATE	5322	12	2	7.37
HDDT GERMAN-NUMER	1000	24	2	2.33
HDDT HEART-V	200	13	2	2.92
HDDT HYPO	3163	25	2	19.95
HDDT ISM	11180	6	2	42.00
HDDT LETTER	20000	16	2	24.35
HDDT OIL	937	49	2	21.85
HDDT PAGE	5473	10	2	8.77
HDDT PENDIGITS	10992	16	2	8.63
HDDT PHONEME	5404	5	2	2.41
HDDT PHOSS	11411	480	2	17.62
HDDT SATIMAGE	6430	36	2	9.29
HDDT SEGMENT	2310	19	2	6.00

Table S1: Properties of HDDT data sets used in experiments.

Data set	Instances	Attributes	Classes	Imbalance ratio
KEEL abalone19	4174	8	2	129.44
KEEL abalone9-18	731	8	2	16.40
KEEL cleveland-0_vs_4	177	13	2	12.62
KEEL ecoli-0-1-3-7_vs_2-6	281	7	2	39.14
KEEL ecoli-0-1-4-6_vs_5	280	6	2	13.00
KEEL ecoli-0-1-4-7_vs_2-3-5-6	336	7	2	10.59
KEEL ecoli-0-1-4-7_vs_5-6	332	6	2	12.28
KEEL ecoli-0-1_vs_2-3-5	244	7	2	9.17
KEEL ecoli-0-1_vs_5	240	6	2	11.00
KEEL ecoli-0-2-3-4_vs_5	202	7	2	9.10
KEEL ecoli-0-2-6-7_vs_3-5	224	7	2	9.18
KEEL ecoli-0-3-4-6_vs_5	205	7	2	9.25
KEEL ecoli-0-3-4-7_vs_5-6	257	7	2	9.28
KEEL ecoli-0-3-4_vs_5	200	7	2	9.00
KEEL ecoli-0-4-6_vs_5	203	6	2	9.15
KEEL ecoli-0-6-7_vs_3-5	222	7	2	9.09
KEEL ecoli-0-6-7_vs_5	220	6	2	10.00
KEEL ecoli-0_vs_1	220	7	2	1.86
KEEL ecoli1	336	7	2	3.36
KEEL ecoli2	336	7	2	5.46
KEEL ecoli3	336	7	2	8.60
KEEL glass1	214	9	2	1.82
KEEL glass2	214	9	2	11.59
KEEL glass4	214	9	2	15.46
KEEL glass5	214	9	2	22.78
KEEL glass6	214	9	2	6.38
KEEL haberman	306	3	2	2.78
KEEL iris0	150	4	2	2.00
KEEL led7digit-0-2-4-5-6-7-8-9_vs_1	443	7	2	10.97
KEEL new-thyroid1	215	5	2	5.14
KEEL new-thyroid2	215	5	2	5.14
KEEL page-blocks-1-3_vs_4	472	10	2	15.86
KEEL pima	768	8	2	1.87
KEEL shuttle-c0-vs-c4	1829	9	2	13.87
KEEL shuttle-c2-vs-c4	129	9	2	20.50
KEEL vehicle0	846	18	2	3.25
KEEL vehicle1	846	18	2	2.90
KEEL vehicle2	846	18	2	2.88
KEEL vehicle3	846	18	2	2.99
KEEL vowel0	988	13	2	9.98
KEEL wisconsin	683	9	2	1.86
KEEL yeast-0-2-5-6_vs_3-7-8-9	1004	8	2	9.14
KEEL yeast-0-2-5-7-9_vs_3-6-8	1004	8	2	9.14
KEEL yeast-0-3-5-9_vs_7-8	506	8	2	9.12
KEEL yeast-0-5-6-7-9_vs_4	528	8	2	9.35
KEEL yeast-1-2-8-9_vs_7	947	8	2	30.57
KEEL yeast-1-4-5-8_vs_7	693	8	2	22.10
KEEL yeast-1_vs_7	459	7	2	14.30
KEEL yeast-2_vs_4	514	8	2	9.08
KEEL yeast-2_vs_8	482	8	2	23.10
KEEL yeast1	1484	8	2	2.46
KEEL yeast3	1484	8	2	8.10
KEEL yeast4	1484	8	2	28.10
KEEL yeast5	1484	8	2	32.73
KEEL yeast6	1484	8	2	41.40

Table S2: Properties of KEEL data sets used in experiments.

Data set	Instances	Attributes	Classes	Source
Adenocarcinoma	76	9868	2	Ramaswamy et al. (2003)
Breast2	77	4769	2	van 't Veer et al. (2002)
Breast3	95	4869	2	van 't Veer et al. (2002)
Colon	62	2001	2	Alon et al. (1999)
Leukemia	38	3052	2	Golub (1999)
Lymphoma	62	4026	2	Alizadeh et al. (2000)
Prostate	38	3052	2	Singh et al. (2002)

Table S3: Properties of microarray data sets used in experiments.

	Instances	Attributes	Classes
UCI iris	150	4	3
UCI car	1728	6	4
UCI monks-1	556	6	2
UCI wine	178	13	3
UCI breast-cancer-wisconsin (bcw)	683	9	2
UCI glass	214	9	7

Table S4: Basic properties of small UCI data sets.

	Instances	Attributes	Classes
UCI led	100000	7	9
UCI waveform	100000	21	3
UCI adult	32561	14	2

Table S5: Basic properties of large UCI data sets used in data pruning benchmark.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750.

Golub, T. R. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537.

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1):49–54.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.