Ensembles of Nested Dichotomies with Multiple Subset Evaluation

Supplementary Materials

Tim Leathart ✉, Eibe Frank, Bernhard Pfahringer and Geoffrey Holmes

Department of Computer Science, University of Waikato, New Zealand
tml15@students.waikato.ac.nz, {eibe,bernhard,geoff}@waikato.ac.nz

A Full Results Tables

In our results tables, a filled circle (•) indicates a statistically significant RMSE reduction, and an open circle (○) indicates a statistically significant RMSE gain ($p = 0.05$) by using a particular value of λ compared to $\lambda = 1$. To establish significance, we used the corrected resampled paired t-test [1] in all of our experiments. Note that the *corrected* version of the paired t-test has been shown to have Type I error at the significance level. The best result for each row, regardless of statistical significance, is in bold face. The datasets used in the experiments and their characteristics are listed in Table 1.

Tables 2 and 3 show the RMSE of individual NDs built with class-balanced and random-pair selection, and compares the RMSE for $\lambda \in \{3, 5, 7\}$ to that of $\lambda = 1$. For class-balanced selection, there is a universal reduction in RMSE for all datasets as λ increases, with the majority of these reductions being statistically significant (compared to $\lambda = 1$). The top-performing model is $\lambda = 7$ in all cases except for `audiology` and `yeast`. There are no datasets for which predictive performance is degraded by adopting multiple subset evaluation. In the case of random-pair selection, the results are less homogeneous; however, there are still many datasets with statistically significant reductions in RMSE. There are two datasets for which we see an increase in RMSE, but these are not statistically significant at $p = 0.05$.

Tables 4 and 5 show the RMSE of bagged ensembles of ten NDs built with class-balanced and random-pair selection, considering values of $\lambda \in \{1, 3, 5, 7\}$. When multiple subset evaluation is employed, there is a statistically significant RMSE reduction in almost all datasets when class-balanced selection is used, and several datasets also see significant improvements when random-pair selection is used. There is one dataset (`mfeat-fourier`) for which RMSE is significantly greater.

Tables 6 and 7 show the RMSE of AdaBoost ensembles of ten NDs built with class-balanced and random-pair selection, and compare the RMSE for the same values of λ. For class-balanced selection, RMSE is generally improved with several datasets showing statistically significant reductions. There are two datasets for which multiple subset evaluation increases the RMSE, and for one of these (`letter`), the increase is significant. Results are less consistent for random-pair

Table 1: The datasets used in our experiments.

Dataset	**Classes**	**Instances**	**Features**
audiology	24	226	70
krkopt	18	28056	7
LED24	10	5000	25
letter	26	20000	17
mfeat-factors	10	2000	217
mfeat-fourier	10	2000	77
mfeat-karhunen	10	2000	65
mfeat-morph	10	2000	7
mfeat-pixel	10	2000	241
MNIST	10	70000	784
optdigits	10	5620	65
page-blocks	5	5473	11
pendigits	10	10992	17
segment	7	2310	20
usps	10	9298	257
vowel	11	990	14
yeast	10	1484	9

selection, with few significant results in either direction. This is reflected in the critical differences plot (in the main paper), which shows single subset evaluation to be statistically equivalent to multiple subset selection for all values of λ, with $\lambda = 7$ performing markedly worse on average.

References

1. Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning **52**(3), 239–281 (2003)

Table 2: RMSE of individual class-balanced NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1541 ± 0.03	**0.1462 ± 0.03**	0.1503 ± 0.02	0.1492 ± 0.03
krkopt	0.2128 ± 0.00	0.2123 ± 0.00	0.2117 ± 0.00 •	**0.2116 ± 0.00** •
led24	0.2200 ± 0.01	0.2101 ± 0.01 •	0.2066 ± 0.01 •	**0.2050 ± 0.01** •
letter	0.1603 ± 0.00	0.1534 ± 0.00 •	0.1500 ± 0.00 •	**0.1482 ± 0.00** •
mfeat-factors	0.1219 ± 0.02	0.1109 ± 0.02	0.1094 ± 0.02	**0.1066 ± 0.02**
mfeat-fourier	0.1997 ± 0.01	0.1967 ± 0.01	0.1960 ± 0.01	**0.1957 ± 0.01**
mfeat-karhunen	0.1543 ± 0.02	0.1397 ± 0.02 •	0.1388 ± 0.02 •	**0.1366 ± 0.02** •
mfeat-morph	0.2204 ± 0.02	0.2041 ± 0.01 •	0.2004 ± 0.01 •	**0.1956 ± 0.01** •
mfeat-pixel	0.1531 ± 0.02	0.1368 ± 0.02 •	0.1309 ± 0.02 •	**0.1291 ± 0.02** •
MNIST	0.1540 ± 0.01	0.1422 ± 0.01 •	0.1377 ± 0.01 •	**0.1358 ± 0.01** •
optdigits	0.1354 ± 0.02	0.1224 ± 0.01 •	0.1169 ± 0.01 •	**0.1148 ± 0.01** •
page-blocks	0.1210 ± 0.01	0.1130 ± 0.01 •	0.1120 ± 0.01 •	**0.1109 ± 0.01** •
pendigits	0.1622 ± 0.02	0.1405 ± 0.02 •	0.1339 ± 0.01 •	**0.1298 ± 0.01** •
segment	0.1599 ± 0.03	0.1354 ± 0.02 •	0.1245 ± 0.02 •	**0.1183 ± 0.02** •
usps	0.1407 ± 0.01	0.1275 ± 0.01 •	0.1249 ± 0.01 •	**0.1232 ± 0.01** •
vowel	0.2382 ± 0.01	0.2205 ± 0.02 •	0.2101 ± 0.02 •	**0.2045 ± 0.02** •
yeast	0.2401 ± 0.01	0.2392 ± 0.01	**0.2378 ± 0.01**	0.2378 ± 0.01

Table 3: RMSE of individual random-pair NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1365 ± 0.02	0.1369 ± 0.03	**0.1320 ± 0.02**	0.1356 ± 0.02
krkopt	0.2094 ± 0.00	0.2092 ± 0.00	0.2090 ± 0.00 •	**0.2090 ± 0.00**
led24	0.2000 ± 0.01	0.1981 ± 0.01	0.1977 ± 0.01 •	**0.1973 ± 0.01** •
letter	0.1327 ± 0.00	0.1272 ± 0.00 •	0.1257 ± 0.00 •	**0.1246 ± 0.00** •
mfeat-factors	0.0923 ± 0.02	0.0860 ± 0.02	0.0855 ± 0.02	**0.0839 ± 0.01**
mfeat-fourier	**0.1936 ± 0.01**	0.1963 ± 0.01	0.1991 ± 0.01	0.2031 ± 0.01
mfeat-karhunen	**0.1333 ± 0.02**	0.1404 ± 0.02	0.1425 ± 0.02	0.1422 ± 0.01
mfeat-morph	0.1858 ± 0.01	0.1837 ± 0.01	**0.1827 ± 0.01**	0.1828 ± 0.01
mfeat-pixel	0.1274 ± 0.02	0.1178 ± 0.02	0.1183 ± 0.02	**0.1168 ± 0.02**
MNIST	0.1285 ± 0.01	0.1192 ± 0.00 •	0.1165 ± 0.00 •	**0.1159 ± 0.00** •
optdigits	0.1080 ± 0.01	0.1005 ± 0.01	**0.0986 ± 0.01**	0.0990 ± 0.01
page-blocks	0.1070 ± 0.01	0.1037 ± 0.01	**0.1033 ± 0.01**	0.1033 ± 0.01
pendigits	0.1204 ± 0.01	0.1036 ± 0.01 •	0.0993 ± 0.01 •	**0.0963 ± 0.01** •
segment	0.1109 ± 0.02	0.1017 ± 0.01	**0.0996 ± 0.01**	0.0999 ± 0.01
usps	0.1152 ± 0.01	0.1101 ± 0.01	**0.1089 ± 0.01**	0.1090 ± 0.00
vowel	0.1600 ± 0.02	0.1540 ± 0.02	0.1531 ± 0.02	**0.1527 ± 0.02**
yeast	0.2354 ± 0.01	0.2347 ± 0.01	0.2348 ± 0.01	**0.2348 ± 0.01**

Table 4: RMSE of an ensemble of 10 bagged class-balanced NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1151 ± 0.01	0.1137 ± 0.01	0.1126 ± 0.01	**0.1120 ± 0.01**
krkopt	0.2112 ± 0.00	0.2104 ± 0.00 •	0.2101 ± 0.00 •	**0.2101 ± 0.00** •
led24	0.2070 ± 0.00	0.1999 ± 0.00 •	0.1983 ± 0.00 •	**0.1977 ± 0.00** •
letter	0.1489 ± 0.00	0.1400 ± 0.00 •	0.1366 ± 0.00 •	**0.1345 ± 0.00** •
mfeat-factors	0.0763 ± 0.01	0.0707 ± 0.01 •	0.0694 ± 0.01 •	**0.0684 ± 0.01** •
mfeat-fourier	0.1663 ± 0.01	0.1619 ± 0.01 •	0.1607 ± 0.01 •	**0.1598 ± 0.01** •
mfeat-karhunen	0.1098 ± 0.01	0.0981 ± 0.01 •	0.0947 ± 0.01 •	**0.0936 ± 0.01** •
mfeat-morph	0.2034 ± 0.01	0.1911 ± 0.01 •	0.1875 ± 0.01 •	**0.1859 ± 0.01** •
mfeat-pixel	0.0980 ± 0.01	0.0910 ± 0.01 •	0.0889 ± 0.01 •	**0.0883 ± 0.01** •
MNIST	0.1281 ± 0.00	0.1181 ± 0.00 •	0.1156 ± 0.00 •	**0.1143 ± 0.00** •
optdigits	0.0974 ± 0.01	0.0850 ± 0.00 •	0.0818 ± 0.00 •	**0.0803 ± 0.00** •
page-blocks	0.1114 ± 0.01	0.1067 ± 0.01 •	0.1055 ± 0.01 •	**0.1054 ± 0.01** •
pendigits	0.1259 ± 0.01	0.1060 ± 0.00 •	0.1002 ± 0.00 •	**0.0970 ± 0.00** •
segment	0.1283 ± 0.01	0.1090 ± 0.01 •	0.1027 ± 0.01 •	**0.1007 ± 0.01** •
usps	0.1070 ± 0.00	0.0984 ± 0.00 •	0.0957 ± 0.00 •	**0.0947 ± 0.00** •
vowel	0.2102 ± 0.01	0.1851 ± 0.01 •	0.1734 ± 0.01 •	**0.1655 ± 0.01** •
yeast	0.2361 ± 0.00	0.2355 ± 0.01	0.2348 ± 0.01 •	**0.2346 ± 0.01** •

Table 5: RMSE of an ensemble of 10 bagged random-pair NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1082 ± 0.01	**0.1073 ± 0.01**	0.1083 ± 0.02	0.1083 ± 0.02
krkopt	0.2088 ± 0.00	0.2088 ± 0.00	0.2088 ± 0.00	**0.2088 ± 0.00**
led24	0.1963 ± 0.01	0.1959 ± 0.01	0.1959 ± 0.01	**0.1958 ± 0.01**
letter	0.1208 ± 0.00	0.1166 ± 0.00 •	0.1155 ± 0.00 •	**0.1151 ± 0.00** •
mfeat-factors	0.0629 ± 0.01	0.0594 ± 0.01 •	0.0584 ± 0.01 •	**0.0583 ± 0.01** •
mfeat-fourier	**0.1596 ± 0.01**	0.1612 ± 0.01	0.1628 ± 0.01 ◦	0.1635 ± 0.01 ◦
mfeat-karhunen	0.0924 ± 0.01	0.0924 ± 0.01	**0.0920 ± 0.01**	0.0923 ± 0.01
mfeat-morph	0.1817 ± 0.01	0.1815 ± 0.01	**0.1814 ± 0.01**	0.1815 ± 0.01
mfeat-pixel	0.0913 ± 0.01	0.0883 ± 0.01	**0.0876 ± 0.01** •	0.0876 ± 0.01 •
MNIST	0.1099 ± 0.00	**0.1080 ± 0.00** •	0.1081 ± 0.00 •	0.1084 ± 0.00 •
optdigits	0.0762 ± 0.00	**0.0717 ± 0.01** •	0.0726 ± 0.01 •	0.0731 ± 0.01 •
page-blocks	0.1023 ± 0.01	**0.1006 ± 0.01** •	0.1008 ± 0.01	0.1008 ± 0.01
pendigits	0.0938 ± 0.00	0.0838 ± 0.01 •	0.0807 ± 0.00 •	**0.0786 ± 0.00** •
segment	0.0984 ± 0.01	0.0934 ± 0.01 •	**0.0928 ± 0.01** •	0.0931 ± 0.01 •
usps	0.0920 ± 0.00	**0.0910 ± 0.00**	0.0919 ± 0.00	0.0921 ± 0.00
vowel	0.1294 ± 0.01	**0.1263 ± 0.01**	0.1265 ± 0.01	0.1270 ± 0.01
yeast	0.2336 ± 0.01	**0.2336 ± 0.01**	0.2336 ± 0.01	0.2337 ± 0.01

Table 6: RMSE of an ensemble of 10 boosted class-balanced NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1152 ± 0.02	0.1144 ± 0.02	**0.1138 ± 0.02**	0.1146 ± 0.02
krkopt	0.2130 ± 0.00	0.2123 ± 0.00	0.2121 ± 0.00 •	**0.2119 ± 0.00** •
led24	0.2617 ± 0.01	0.2588 ± 0.00	0.2581 ± 0.00	**0.2580 ± 0.00** •
letter	**0.1619 ± 0.01**	0.1830 ± 0.01 ○	0.1872 ± 0.01 ○	0.1862 ± 0.00 ○
mfeat-factors	0.0698 ± 0.02	0.0646 ± 0.02	0.0650 ± 0.02	**0.0637 ± 0.02**
mfeat-fourier	0.1714 ± 0.01	0.1718 ± 0.01	**0.1712 ± 0.01**	0.1716 ± 0.01
mfeat-karhunen	0.0997 ± 0.01	0.0943 ± 0.01	0.0955 ± 0.02	**0.0934 ± 0.01**
mfeat-morph	0.2558 ± 0.02	0.2488 ± 0.01	0.2464 ± 0.01	**0.2463 ± 0.01**
mfeat-pixel	0.1003 ± 0.01	0.0963 ± 0.01	0.0958 ± 0.01	**0.0948 ± 0.01**
MNIST	0.1274 ± 0.00	0.1223 ± 0.00 •	0.1221 ± 0.00 •	**0.1216 ± 0.00** •
optdigits	0.0765 ± 0.01	0.0720 ± 0.01	0.0728 ± 0.01	**0.0714 ± 0.01**
page-blocks	0.1208 ± 0.01	0.1179 ± 0.01	**0.1175 ± 0.01**	0.1185 ± 0.01
pendigits	0.1015 ± 0.01	**0.0891 ± 0.01** •	0.0894 ± 0.01 •	0.0894 ± 0.01 •
segment	0.1146 ± 0.02	**0.1115 ± 0.02**	0.1134 ± 0.01	0.1124 ± 0.02
usps	0.1028 ± 0.01	0.0994 ± 0.01	0.0990 ± 0.01 •	**0.0985 ± 0.01** •
vowel	0.2559 ± 0.02	0.1921 ± 0.04 •	0.1509 ± 0.02 •	**0.1359 ± 0.02** •
yeast	**0.2819 ± 0.02**	0.2871 ± 0.01	0.2830 ± 0.02	0.2864 ± 0.02

Table 7: RMSE of an ensemble of 10 boosted random-pair NDs for the range of values of λ.

Dataset	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$	$\lambda = 7$
audiology	0.1127 ± 0.02	0.1137 ± 0.02	0.1144 ± 0.02	**0.1117 ± 0.02**
krkopt	0.2095 ± 0.00	0.2094 ± 0.00	0.2092 ± 0.00	**0.2092 ± 0.00**
led24	0.2567 ± 0.00	0.2566 ± 0.00	**0.2565 ± 0.00**	0.2567 ± 0.00
letter	0.1784 ± 0.00	0.1766 ± 0.00 •	0.1764 ± 0.00 •	**0.1764 ± 0.00** •
mfeat-factors	0.0618 ± 0.02	**0.0613 ± 0.01**	0.0618 ± 0.02	0.0632 ± 0.01
mfeat-fourier	**0.1723 ± 0.01**	0.1759 ± 0.01	0.1737 ± 0.01	0.1773 ± 0.01
mfeat-karhunen	**0.0937 ± 0.01**	0.0964 ± 0.01	0.0947 ± 0.01	0.0954 ± 0.01
mfeat-morph	0.2412 ± 0.00	**0.2410 ± 0.00**	0.2418 ± 0.00	0.2418 ± 0.00
mfeat-pixel	**0.0936 ± 0.01**	0.0939 ± 0.01	0.0938 ± 0.01	0.0940 ± 0.01
MNIST	**0.1213 ± 0.00**	0.1229 ± 0.00 ○	0.1239 ± 0.00 ○	0.1245 ± 0.00 ○
optdigits	**0.0714 ± 0.01**	0.0718 ± 0.01	0.0726 ± 0.01	0.0740 ± 0.01
page-blocks	0.1163 ± 0.01	0.1147 ± 0.01	**0.1143 ± 0.01**	0.1145 ± 0.01
pendigits	0.0895 ± 0.01	0.0893 ± 0.01	**0.0892 ± 0.01**	0.0897 ± 0.01
segment	0.1069 ± 0.01	0.1064 ± 0.02	**0.1062 ± 0.01**	0.1070 ± 0.02
usps	**0.1005 ± 0.01**	0.1015 ± 0.01	0.1025 ± 0.01	0.1035 ± 0.01
vowel	**0.1161 ± 0.02**	0.1204 ± 0.02	0.1240 ± 0.02	0.1260 ± 0.02 ○
yeast	0.2874 ± 0.01	0.2900 ± 0.01	**0.2874 ± 0.01**	0.2908 ± 0.00