A. Variations on the Column Generation Heuristic

We have explored the following three variations of the heuristic algorithm for column generation described in Section 4:

1. The algorithm can return the best K solutions that it finds instead of a single incumbent solution, potentially reducing the number of CG iterations needed. We have observed however that these solutions tend to correspond to very similar conjunctions and are hence highly correlated. By allowing multiple such columns to enter together, sparsity suffers because the ℓ_1 regularization in (4) has difficulty favoring sparse linear combinations of highly correlated columns over dense ones. For this reason we kept $K = 1$.

2. The algorithm can be generalized to a beam search by considering the children of $B > 1$ parent conjunctions at each degree instead of a single parent. The best B children according to a combination of metrics (10) and (11) are then chosen to become the next parents. To date however, we have not found setting $B > 1$ to be beneficial.

3. The algorithm can be terminated early once a solution with negative objective value is found since any such solution corresponds to a descent direction for problem (7). Termination can be immediate or occur after the current degree. While early termination speeds up each CG iteration, the number of iterations tends to increase because the generated columns are of lower quality.

B. Additional Numerical Results

B.1. Classification

Figures 3–6 show trade-offs between Brier score and weighted rules and between accuracy and weighted rules for all 16 classification datasets.

Tables 5 and 6 show mean test accuracies and corresponding complexities when the methods are optimized for accuracy. For Table 5, the Friedman statistic computed from the mean ranks is 9.74 with a p-value of 0.202, indicating no statistically significant differences in accuracy among the methods. For Table 6, the Friedman statistic is 44.61 (p-value $\sim 10^{-12}$). Post-hoc comparisons with LRRN as the reference show that RuleFit and RuleFitN are significantly more complex at the 0.05 level using Holm's step-down procedure.

B.2. Regression

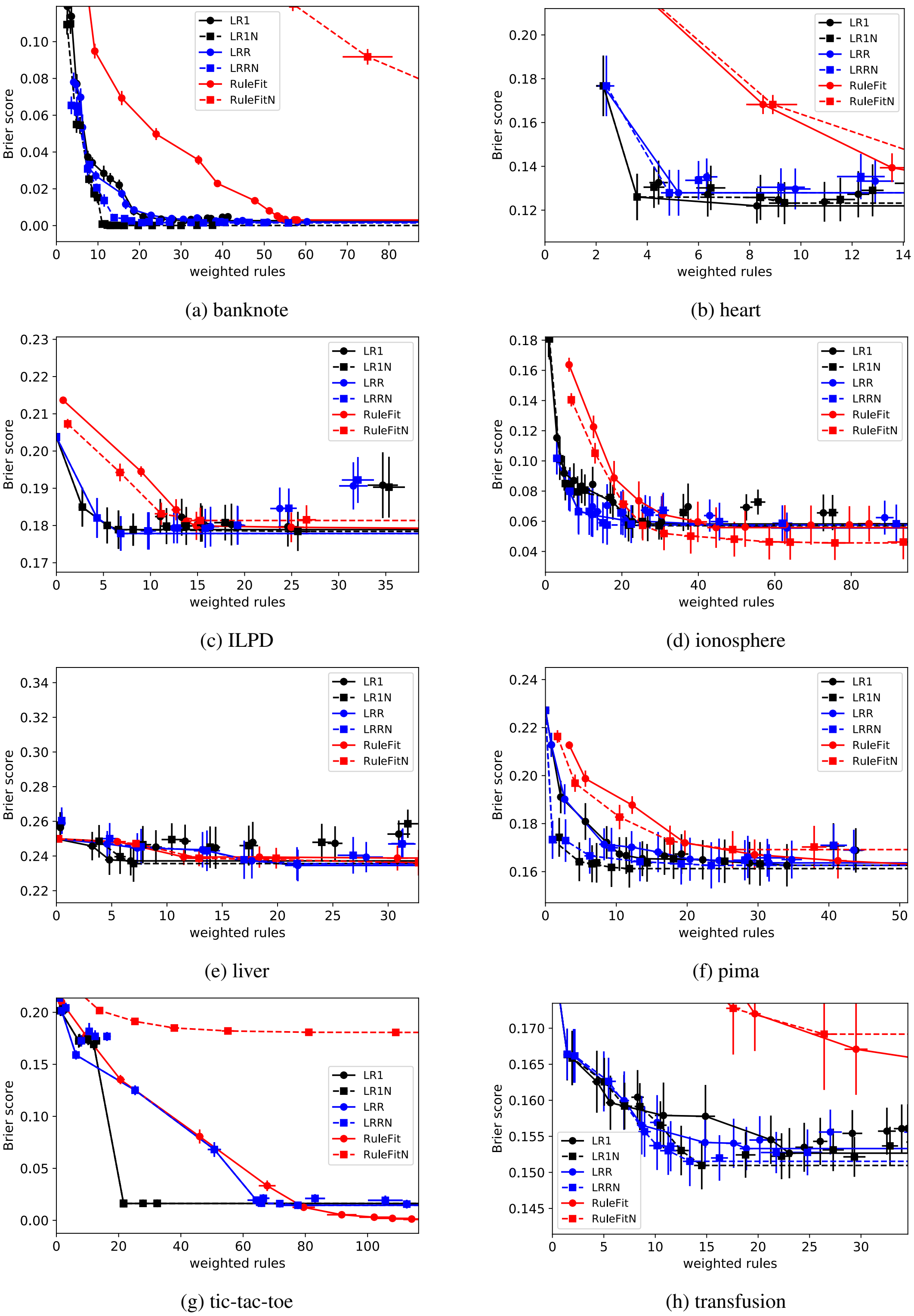
Figure 7 shows the trade-off between R^2 and weighted rules for all 8 regression datasets.

Table 5. Mean test accuracies (%, standard error in parentheses). Best values in **bold**.

dataset	LR1	LRR	RuleFit	LR1N	LRRN	RuleFitN	GBT	SVM
banknote	99.8 (0.1)	99.7 (0.1)	99.6 (0.1)	**100.0** (0.0)	99.9 (0.1)	99.9 (0.1)	99.7 (0.1)	99.9 (0.1)
heart	80.9 (1.6)	84.3 (2.0)	83.3 (1.3)	81.3 (1.8)	**84.6** (1.9)	83.3 (2.0)	82.3 (1.8)	82.6 (1.4)
ILPD	71.0 (1.1)	70.8 (0.5)	71.5 (0.1)	70.6 (0.9)	70.8 (0.7)	71.7 (1.1)	**71.8** (0.2)	71.7 (0.2)
ionosphere	91.2 (1.2)	91.2 (1.3)	93.4 (1.5)	91.7 (1.1)	90.9 (1.6)	94.3 (1.3)	91.2 (1.8)	**94.9** (1.4)
liver	**61.2** (2.0)	59.1 (2.2)	58.0 (2.2)	60.0 (2.6)	58.0 (2.7)	58.6 (2.1)	57.1 (2.5)	58.8 (2.7)
pima	75.5 (1.6)	75.1 (1.4)	75.5 (1.9)	**77.7** (1.3)	75.8 (1.6)	74.7 (1.9)	75.9 (1.9)	77.1 (2.0)
tic-tac-toe	98.3 (0.4)	98.0 (0.6)	**100.0** (0.0)	98.3 (0.4)	98.0 (0.6)	**100.0** (0.0)	99.1 (0.2)	98.3 (0.4)
transfusion	76.7 (0.3)	79.0 (0.9)	75.5 (1.9)	78.7 (0.7)	**79.3** (1.0)	74.7 (1.9)	76.6 (0.3)	76.9 (0.3)
WDBC	97.0 (0.6)	97.9 (0.5)	97.9 (0.4)	97.2 (0.7)	**98.2** (0.4)	96.8 (0.5)	95.6 (0.6)	98.1 (0.4)
adult	84.9 (0.2)	84.9 (0.2)	84.8 (0.2)	85.8 (0.1)	85.9 (0.1)	**87.0** (0.2)	84.8 (0.2)	84.8 (0.1)
bank-mkt	88.7 (0.0)	90.0 (0.1)	88.7 (0.0)	88.7 (0.0)	**90.1** (0.1)	88.7 (0.0)	89.9 (0.1)	88.7 (0.0)
gas	99.5 (0.0)	99.6 (0.1)	99.5 (0.1)	**99.6** (0.0)	99.5 (0.1)	**99.6** (0.0)	99.4 (0.1)	99.5 (0.1)
magic	84.9 (0.3)	85.4 (0.3)	86.7 (0.2)	85.1 (0.3)	85.4 (0.2)	**87.5** (0.2)	87.2 (0.2)	87.4 (0.2)
mushroom	**100.0** (0.0)	**100.0** (0.0)	**100.0** (0.0)	**100.0** (0.0)	**100.0** (0.0)	**100.0** (0.0)	99.9 (0.1)	**100.0** (0.0)
musk	96.8 (0.5)	**98.4** (0.1)	97.6 (0.3)	96.1 (0.7)	98.4 (0.2)	97.8 (0.2)	94.5 (0.5)	97.6 (0.7)
FICO	73.8 (0.3)	73.8 (0.2)	73.8 (0.2)	**74.0** (0.2)	73.9 (0.2)	**74.0** (0.2)	73.3 (0.2)	72.4 (0.4)
mean rank	5.25	4.31	4.84	4.16	3.78	**3.69**	5.75	4.22

Table 6. Mean weighted number of rules (standard error in parentheses) corresponding to Table 5. Best values in **bold**.

dataset	LR1	LRR	RuleFit	LR1N	LRRN	RuleFitN
banknote	32.3 (0.8)	47.2 (3.4)	57.8 (0.7)	**16.4** (0.4)	47.7 (1.6)	1124.9 (67.7)
heart	13.4 (2.7)	5.7 (0.6)	34.3 (0.9)	14.3 (2.1)	**5.2** (0.4)	59.4 (2.4)
ILPD	14.6 (3.7)	38.1 (25.5)	**0.0** (0.0)	14.7 (4.9)	1.9 (1.9)	2106.0 (30.3)
ionosphere	114.4 (28.5)	**85.2** (22.6)	1022.6 (64.9)	130.3 (23.7)	150.7 (49.3)	1225.6 (81.3)
liver	28.9 (5.8)	**20.8** (4.7)	66.7 (11.7)	25.7 (5.7)	34.7 (14.2)	89.7 (36.8)
pima	22.1 (2.3)	27.7 (1.8)	64.8 (1.1)	**12.1** (1.0)	15.5 (2.3)	3211.5 (83.3)
tic-tac-toe	**21.6** (0.0)	67.1 (3.5)	1640.7 (99.2)	**21.6** (0.0)	67.1 (3.5)	1640.7 (99.2)
transfusion	15.2 (4.3)	17.8 (1.1)	64.8 (1.1)	24.3 (2.5)	**11.9** (1.4)	3211.5 (83.3)
WDBC	145.7 (18.0)	283.6 (10.3)	809.4 (89.6)	**86.1** (12.6)	228.4 (30.3)	562.3 (83.3)
adult	87.1 (1.6)	91.5 (4.8)	425.5 (35.0)	**85.8** (2.4)	94.2 (6.2)	719.9 (58.6)
bank-mkt	**0.0** (0.0)	68.6 (9.5)	**0.0** (0.0)	**0.0** (0.0)	83.6 (4.9)	0.2 (0.0)
gas	**483.7** (8.0)	678.2 (17.6)	2663.1 (235.9)	950.2 (12.9)	1259.4 (45.8)	2920.8 (125.7)
magic	**93.1** (2.2)	177.2 (17.5)	496.7 (5.9)	97.9 (2.9)	196.2 (25.0)	1656.0 (11.7)
mushroom	24.7 (0.6)	**18.2** (0.9)	927.9 (58.6)	24.7 (0.6)	**18.2** (0.9)	927.9 (58.6)
musk	**263.0** (39.3)	1002.0 (71.8)	1796.4 (326.1)	313.9 (101.6)	1079.7 (78.4)	2000.3 (314.8)
FICO	92.6 (5.9)	65.9 (3.4)	239.8 (2.5)	81.0 (5.2)	**56.2** (4.1)	183.4 (3.0)
mean rank	**2.25**	2.88	4.75	2.38	3.06	5.69



(a) banknote

(b) heart

(c) ILPD

(d) ionosphere

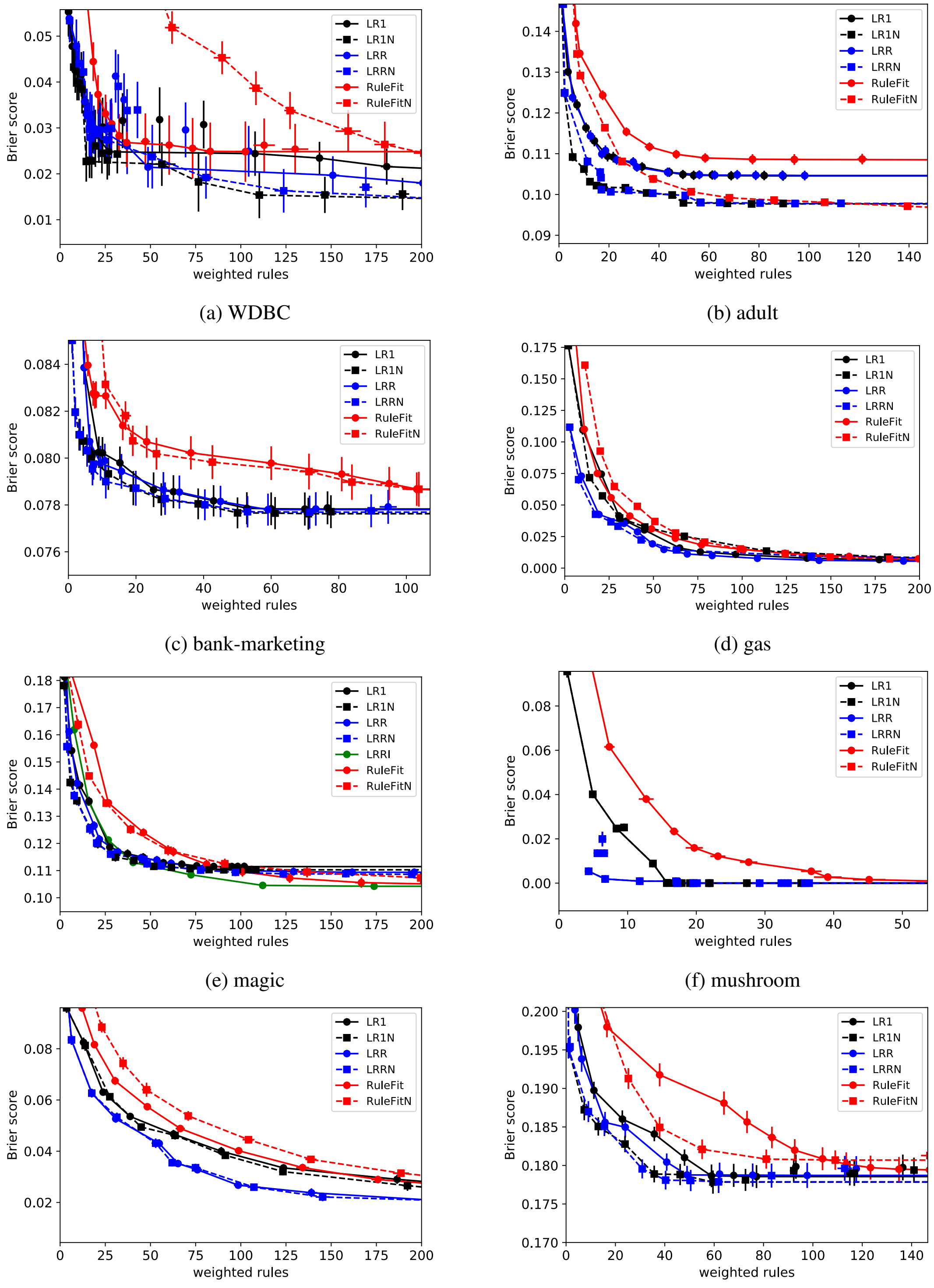
(e) liver

(f) pima

(g) tic-tac-toe

(h) transfusion

Figure 3. Trade-offs between Brier score and weighted number of rules on classification datasets. Pareto efficient points are connected by line segments. Horizontal and vertical bars represent standard errors in the means.



(a) WDBC

(b) adult

(c) bank-marketing

(d) gas

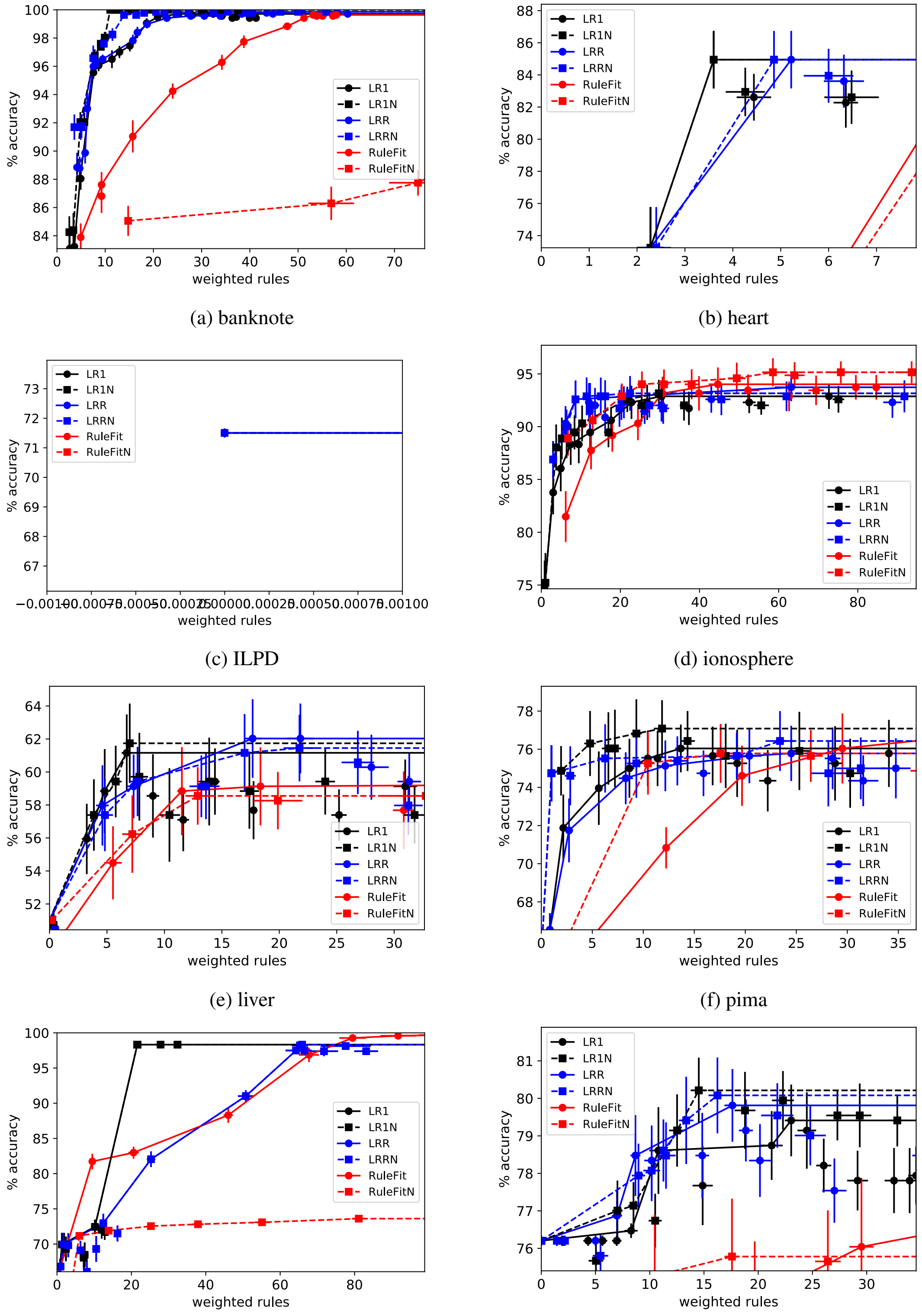
(e) magic

(f) mushroom

(g) musk

(h) FICO

Figure 4. Trade-offs between Brier score and weighted number of rules on classification datasets. Pareto efficient points are connected by line segments. Horizontal and vertical bars represent standard errors in the means.



(a) banknote

(b) heart

(c) ILPD

(d) ionosphere

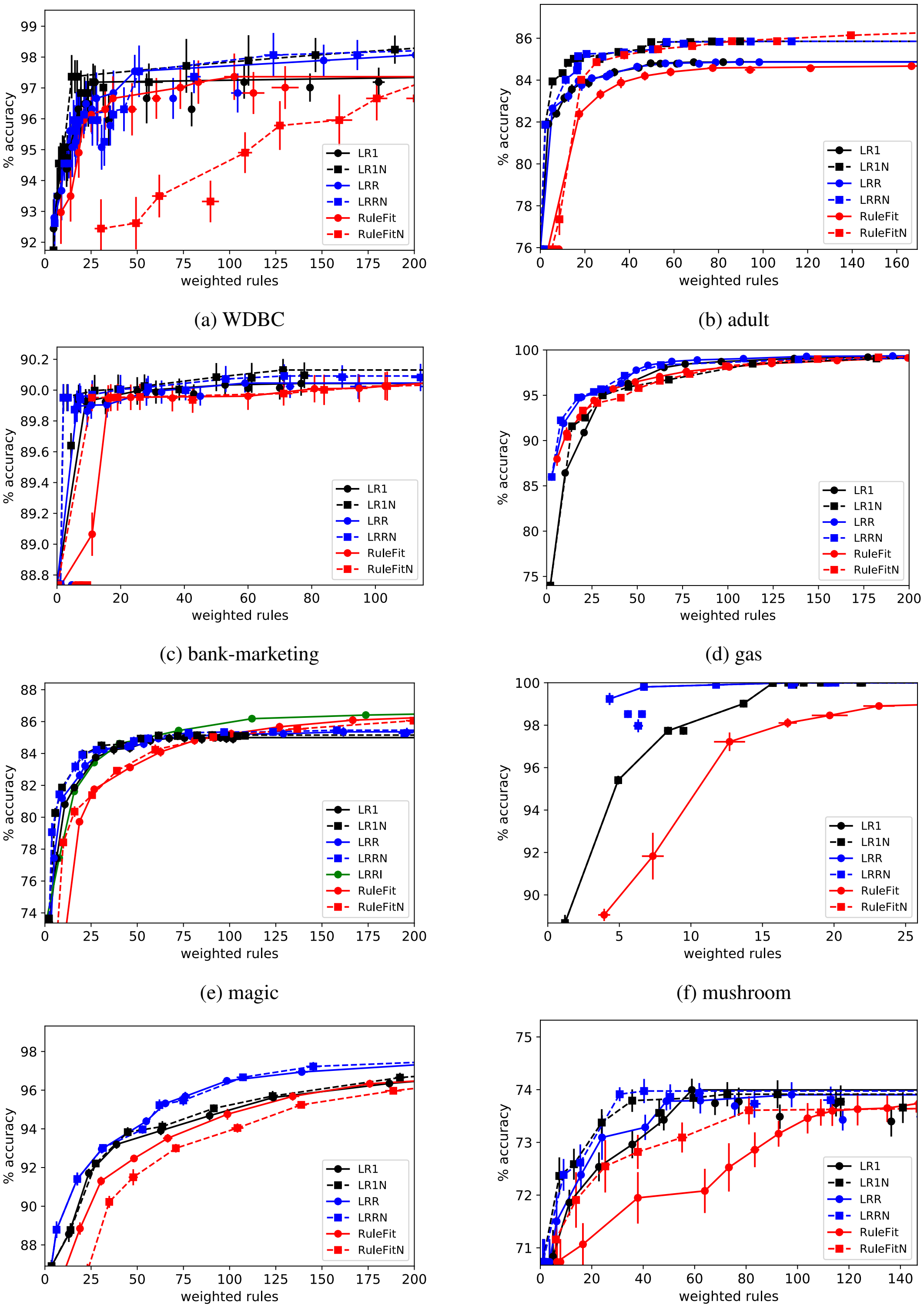
(e) liver

(f) pima

(g) tic-tac-toe

(h) transfusion

Figure 5. Trade-offs between accuracy and weighted number of rules on classification datasets. Pareto efficient points are connected by line segments. Horizontal and vertical bars represent standard errors in the means.



(a) WDBC

(b) adult

(c) bank-marketing

(d) gas

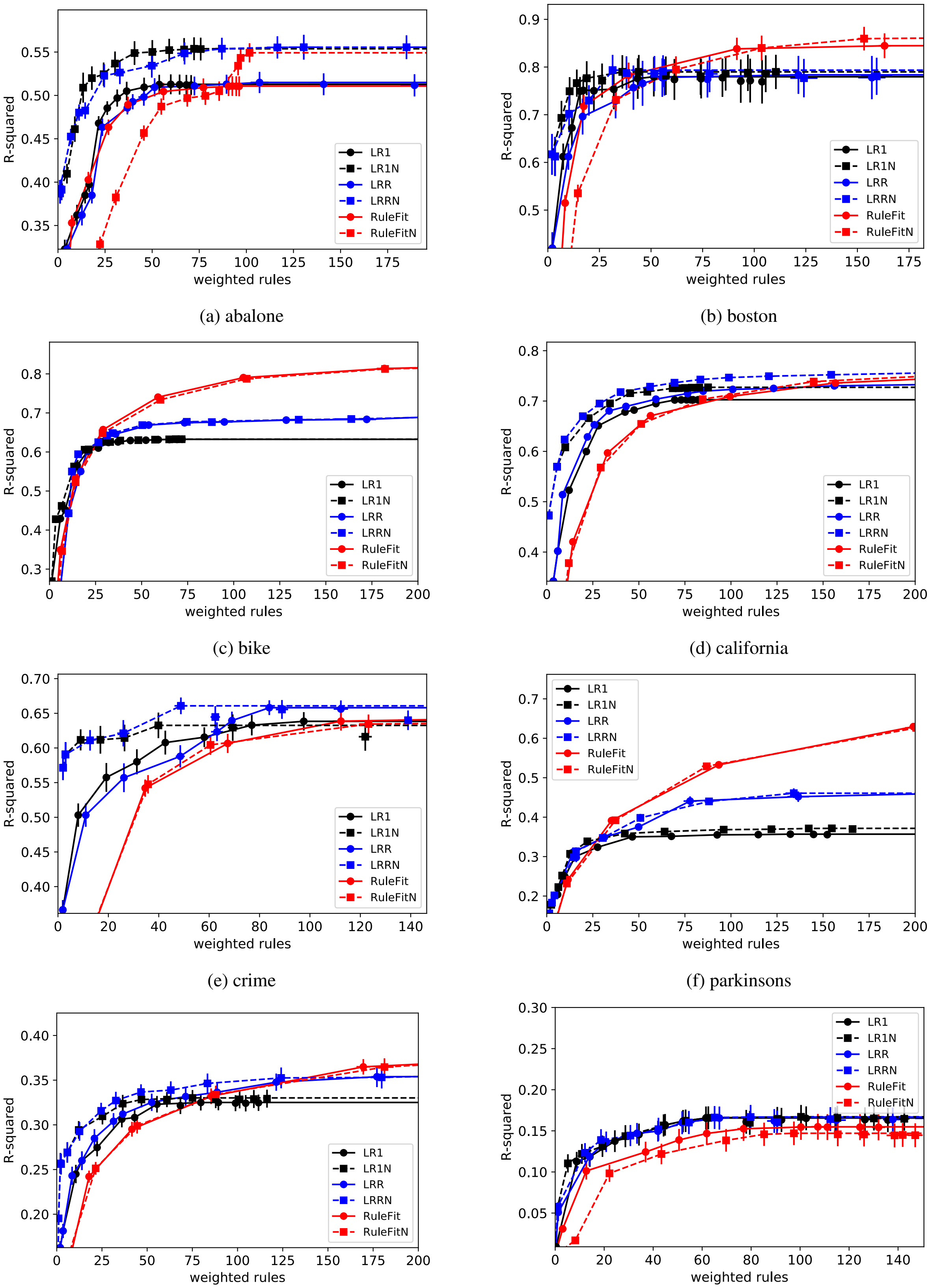
(e) magic

(f) mushroom

(g) musk

(h) FICO

Figure 6. Trade-offs between accuracy and weighted number of rules on classification datasets. Pareto efficient points are connected by line segments. Horizontal and vertical bars represent standard errors in the means.



(a) abalone

(b) boston

(c) bike

(d) california

(e) crime

(f) parkinsons

(g) wine

(h) MEPS

Figure 7. Trade-offs between coefficient of determination R^2 and weighted number of rules on regression datasets. Pareto efficient points are connected by line segments. Horizontal and vertical bars represent standard errors in the means.