Addendum to
Optimistic Active Learning using Mutual Information*

Yuhong Guo and Russ Greiner
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
{ yuhong | greiner }@cs.ualberta.ca

October 21, 2006

Our experimental results are shown in Table 1, whose entries are of the form "**#**MM+M**-wins / #**A**-wins** (s)", where A is the algorithm associated with the column. (The parenthesized number next to each database name is the accuracy one can obtain, using all of the data. We will explain the parenthesized "($\pm\pm$)" value below.) An entry is **bold** if MM+M was statistically better at least 5 more times than it was statistically worse. Each entry in the "**TOTAL** W/L/T" row shows the number of datasets where MM+M "dominated" by this "> 5 measure", when it lost by this same quantity, versus tied. We also note that, wrt these 4 active learners, our MM+M was the best active learner for 11 of the 17 databases (in that it was at least as good as the other 3 approaches, and strictly better than at least one); we marked each such database with a "*".

To describe our second set of evaluations, note that the 100 numbers $\{acc(\ D,\ A[m]\ )\}_{m=1}^{100}$ characterize the performance of A on D. We therefore ran a Wilcoxon signed rank test to determine whether these scores were significantly better for MM+M, or for the "challenger" associated with the column. This produced the first w value in the parentheses "($w\,t$)" in each entry in Table 1, which is "+" if this signed rank test suggests that MM+M is statistically better than the alternative algorithm A at the $p < 0.05$ level, is "–" if this test suggests that A is statistically better at $p < 0.05$, and is "0" otherwise. (Hence, this test suggests that MM+M was statistically better than MU on the AUSTRALIAN database, but it was statistically worse on the BREAST database, and is comparable for CORRAL.) The second value, t, is based instead on a 2-sides t-test, and here again a "+" (resp., "–", "0") means this test shows that MM+M is statistically ($p < 0.05$) better than (resp., worse than, the same as) A.

The numbers in the "Signed Rank Test" (resp., "t-test") row of Table 1 are the totals for the signed rank test (resp, t-test); they are in a similar form as the first evaluation,

*This includes extra data, beyond the material that appears in IJCAI2007.

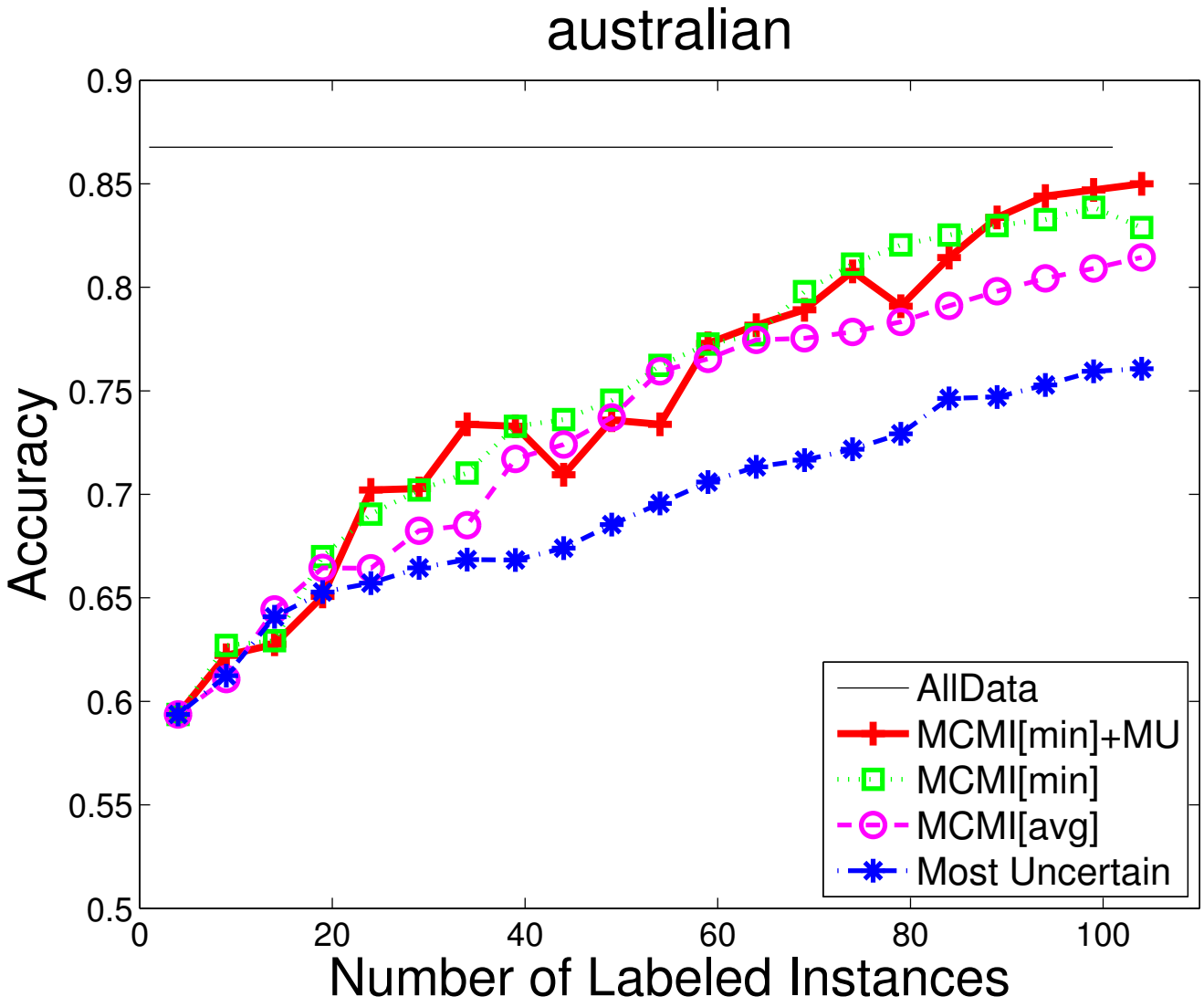
Table 1: Comparing MM+M to other Active Learners.

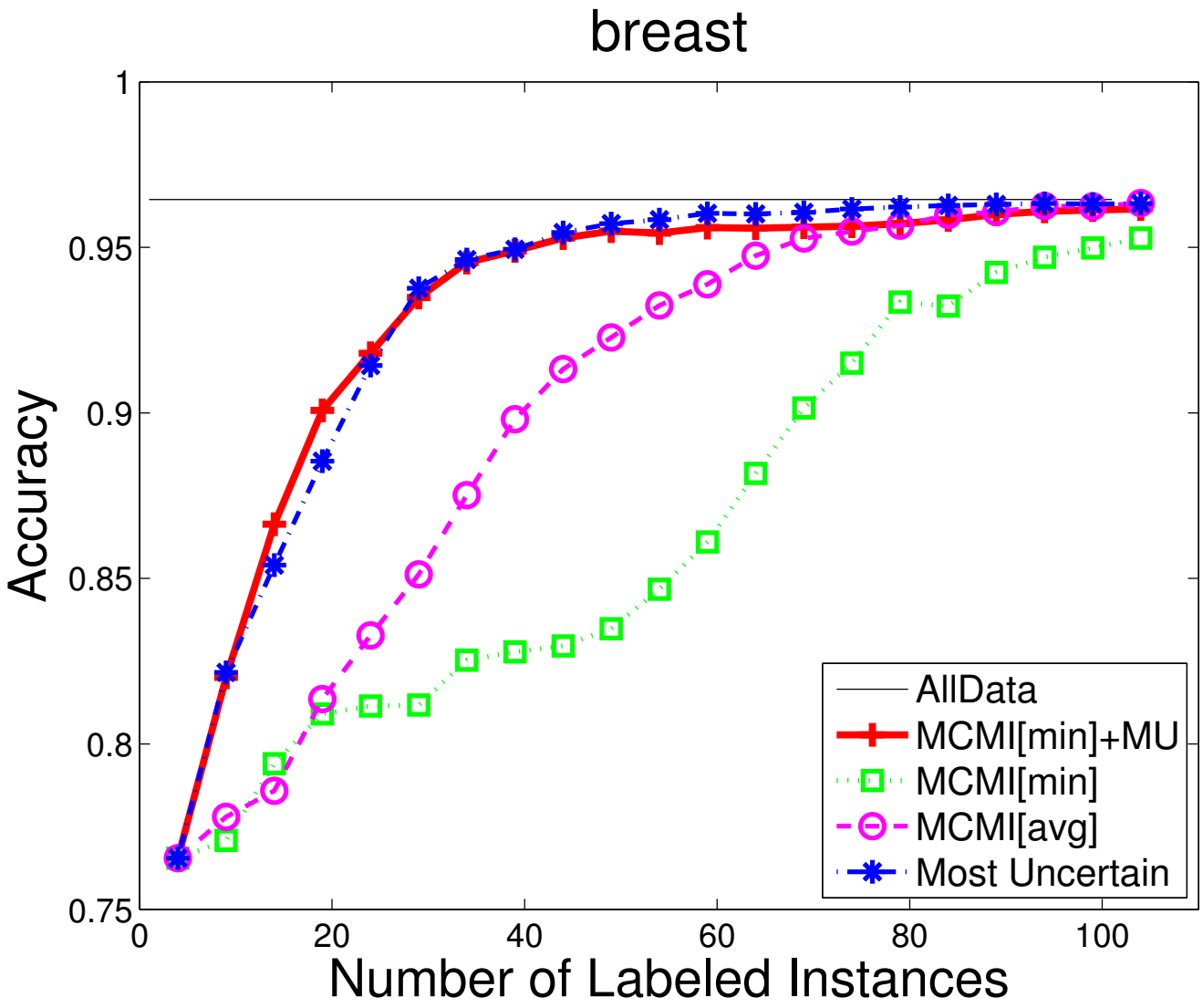
database	vs MU	vs MCMI[min]	vs MCMI[avg]	vs RANDOM	vs MU-SVM
AUSTRALIAN* (0.85)	**58**/ 0 (++)	1 / 0 (00)	**24**/ 0 (++)	11 / 14 (−−)	**16**/ 0 (++)
BREAST* (0.96)	0 / 0 (−−)	**94**/ 0 (++)	**56**/ 0 (++)	**88**/ 0 (++)	0 /53 (−−)
CLEVE* (0.83)	**56**/ 0 (++)	**10**/ 0 (++)	**32**/ 0 (++)	**26**/ 0 (++)	**77**/ 0 (++)
CORRAL (0.91)	5 / 14 (0−)	**44**/ 2 (++)	1 / 0 (00)	**13**/ 0 (++)	0 /23 (−−)
CRX (0.84)	**80**/ 0 (++)	0 / 5 (−−)	**26**/ 0 (++)	2 / 8 (−−)	**13**/ 0 (++)
DIABETES* (0.75)	**87**/ 0 (++)	0 / 1 (00)	**87**/ 1 (++)	11 / 12 (++)	**73**/ 8 (++)
FLARE* (0.83)	**53**/ 0 (++)	**41**/ 0 (++)	**55**/ 0 (++)	**27**/ 2 (++)	**59**/ 1 (++)
GERMAN* (0.71)	**42**/ 0 (++)	**18**/ 0 (++)	**77**/ 0 (++)	0 / 0 (++)	**91**/ 0 (++)
GLASS2* (0.72)	**38**/ 0 (++)	4 / 0 (++)	**24**/ 0 (++)	**8**/ 0 (++)	**17**/ 0 (++)
HEART* (0.84)	**30**/ 0 (++)	**19**/ 0 (++)	**20**/ 0 (++)	0 / 0 (++)	**29**/ 0 (++)
HEPATITIS* (0.88)	**7**/ 0 (++)	0 / 0 (00)	**6**/ 0 (++)	0 / 0 (00)	**50**/ 0 (++)
MOFN (1.0)	0 / 33 (−−)	**93**/ 0 (++)	**67**/ 0 (++)	**31**/ 6 (00)	0 /37 (0−)
PIMA* (0.75)	**85**/ 2 (++)	2 / 0 (++)	**84**/ 2 (++)	0 / 0 (00)	**36**/13 (+0)
VOTE (0.93)	0 / 27 (−−)	**97**/ 0 (++)	0 / 0 (−−)	2 / 0 (++)	0 /16 (−−)
IRIS (0.97)	**68**/ 1 (++)	**17**/ 0 (++)	13 / 23 (00)	**38**/ 1 (++)	- / - ( )
VEHICLE* (0.72)	**56**/ 0 (++)	**84**/ 0 (++)	**38**/ 0 (++)	0 / 19 (−−)	- / - ( )
LYMPHOGRAPHY (0.84)	0 / 4 (0−)	0 / 1 (−−)	0 / 30 (−−)	0 / 2 (−0)	- / - ( )
TOTAL W/L/T	**12** / 3 / 2	**10** / 1 / 6	**13** / 2 / 2	**7** / 2 / 8	**10** / 4 / 0
Signed Rank Test	**12** / 3 / 2	**12** / 2 / 3	**13** / 1 / 3	**10** / 4 / 3	**10** / 3 / 1
t-test	**12** / 5 / 0	**12** / 2 / 3	**13** / 1 / 3	**10** / 3 / 4	**9** / 4 / 1

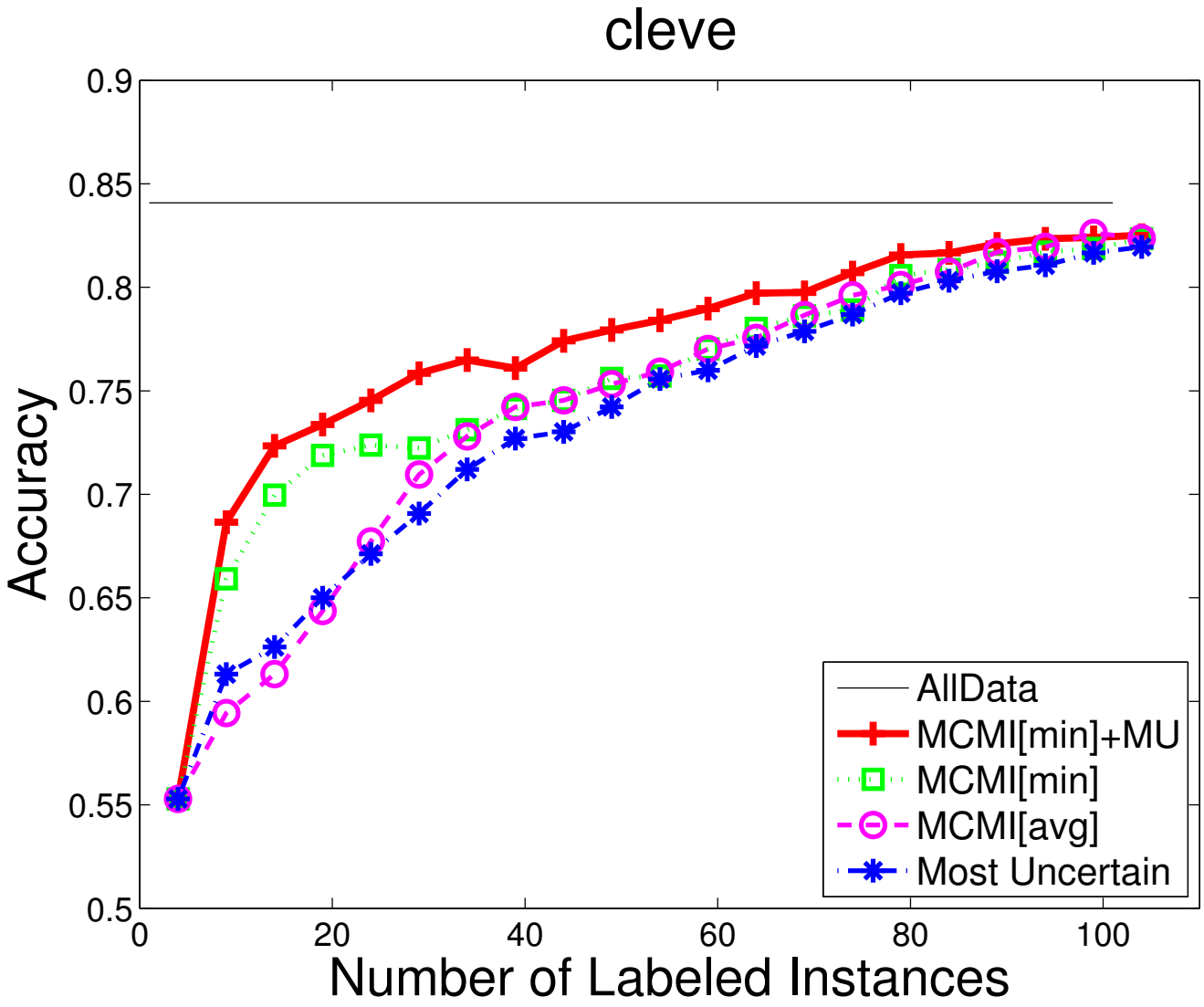
i.e., the number of datasets where MM+M was better/worse/tied. Hence, the Wilcoxon signed rank test shows that MM+M is statistically better than MU for 12 of the 17 datasets, worse on 3, and tied in the remaining 2. Notice the results of these tests (Ranked Sign and t-test) are very similar to our other (ad hoc) scoring measure.

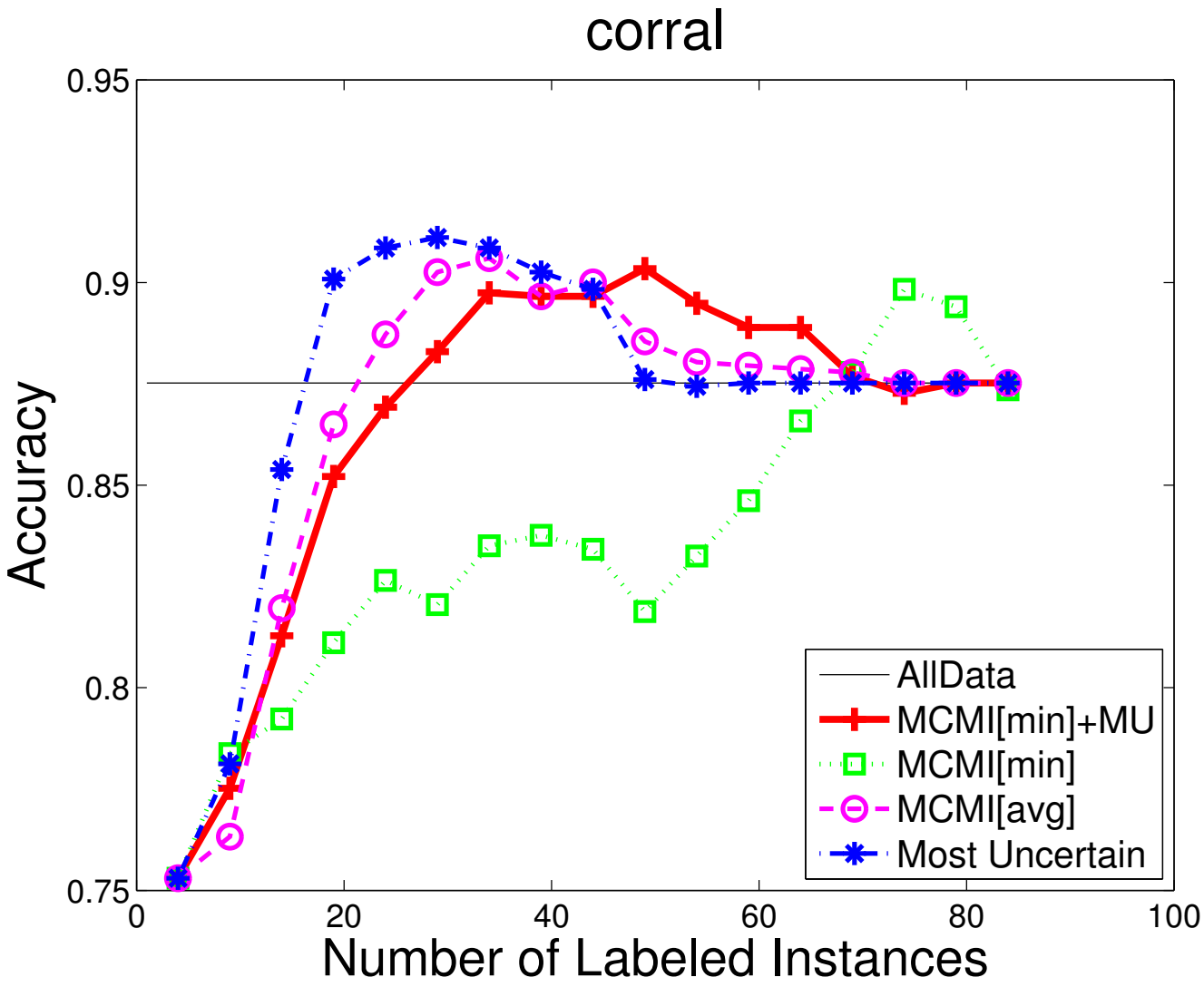
We also considered two other active learners. The "RANDOM" learner simply selected instances at random. The final learner, "MU-SVM", uses a (linear kernel) SVM, and selects the instance closest to the classification boundary — i.e., this is a variant of the "most uncertain" selection criterion. (Note we only consider the 14 binary databases here.) Using the evaluation measures shown above, the results of our experiments appear in the final two columns of Table 1.

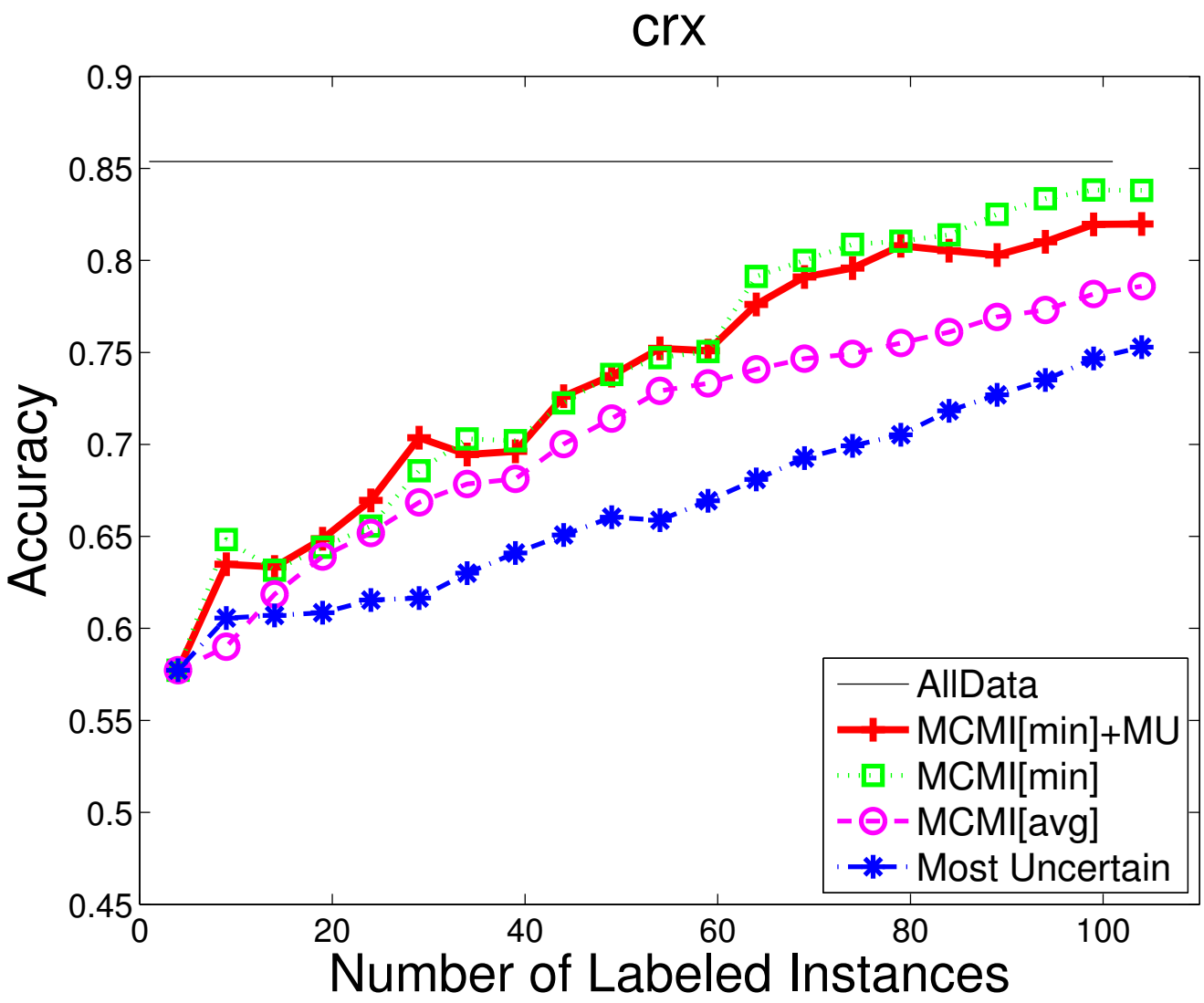
The following graphs show the actual results over all 17 datasets. Each point for each "active learner" is the average over 30 trials. The "AllData" line corresponds to the 5-fold cross validation results, using all of the data.

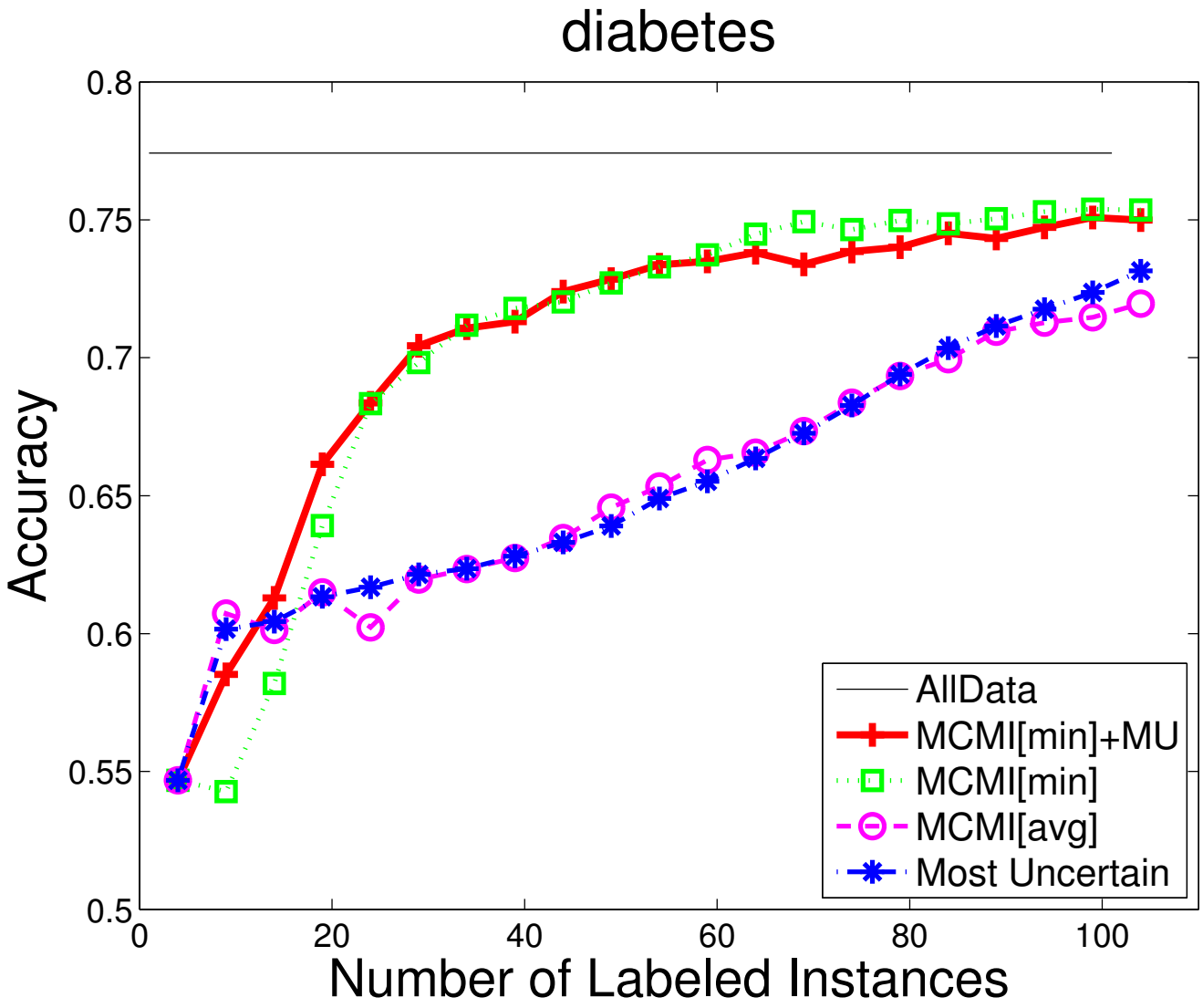

australian
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

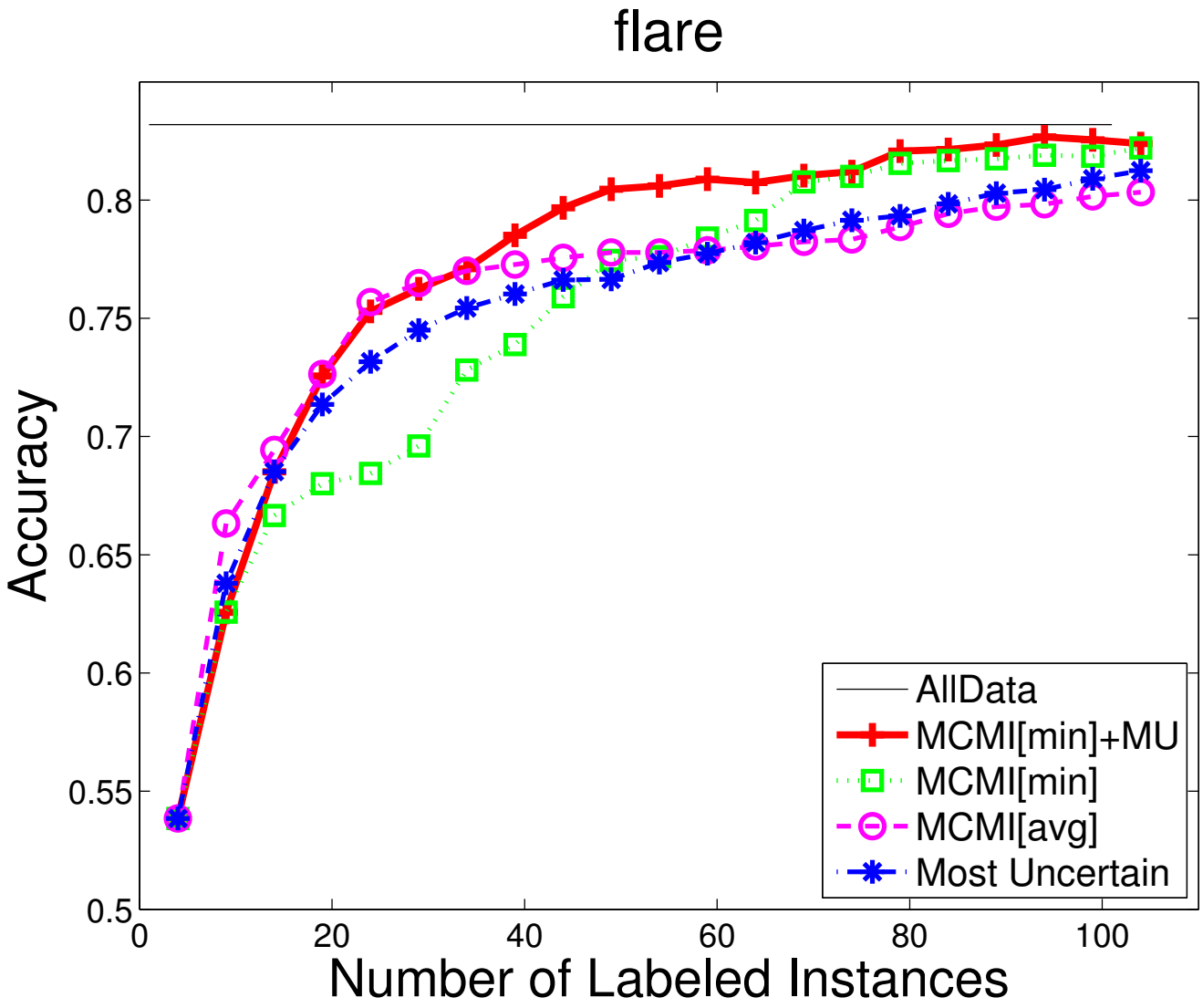

breast
Accuracy
Number of Labeled Instances
1
0.95
0.9
0.85
0.8
0.75
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

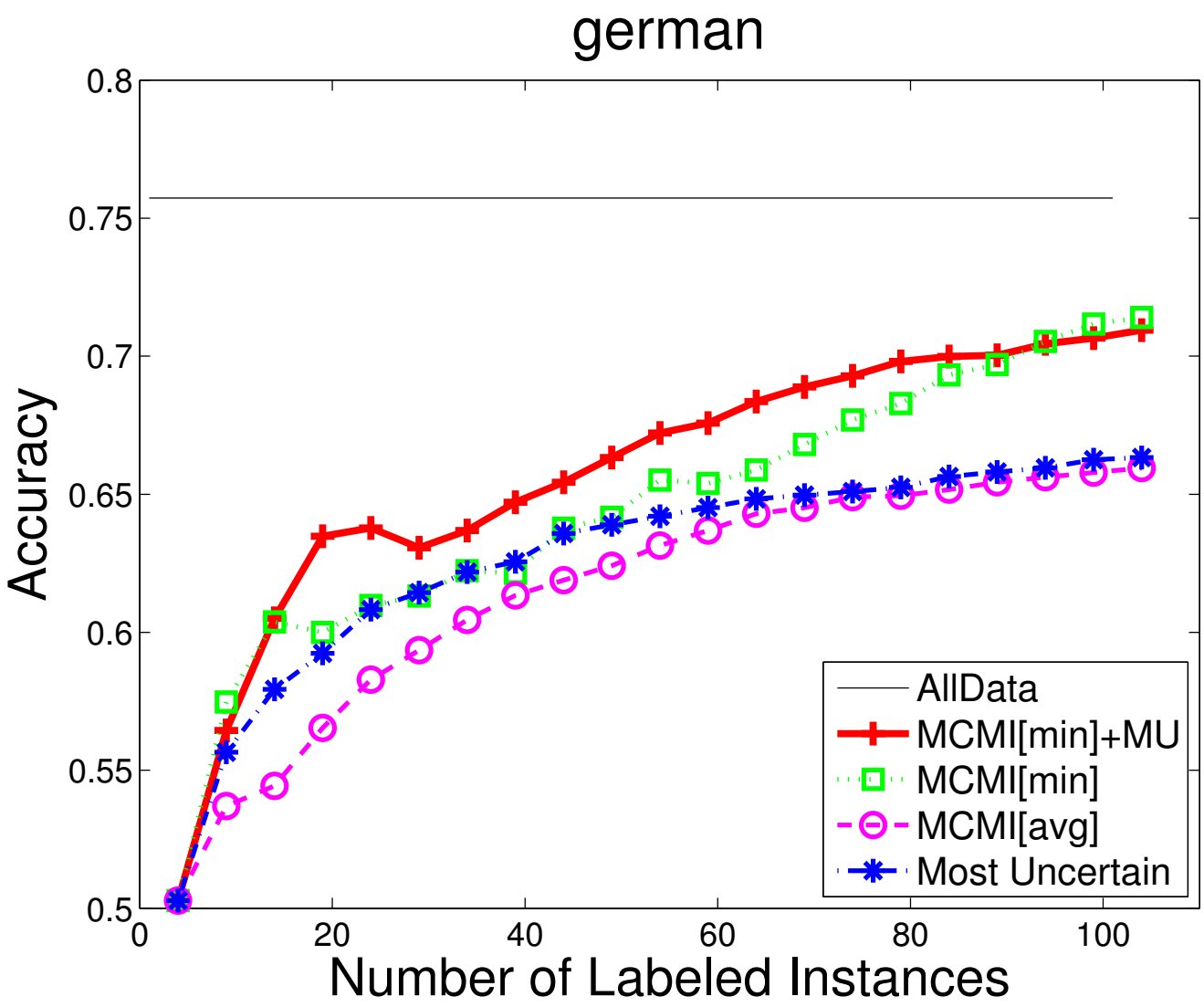

cleve
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

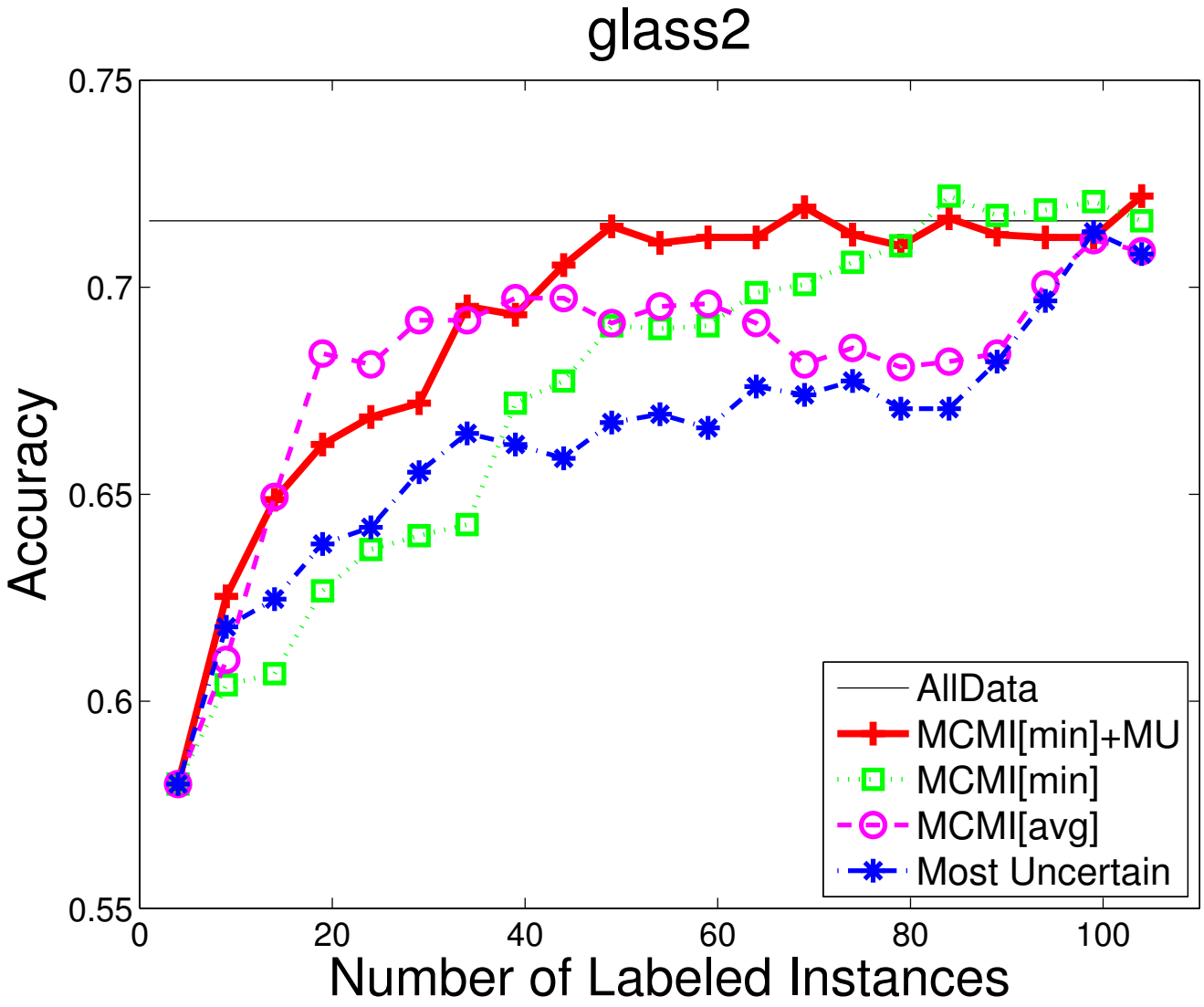

corral
0.95
0.9
0.85
0.8
0.75
Accuracy
0
20
40
60
80
100
Number of Labeled Instances
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

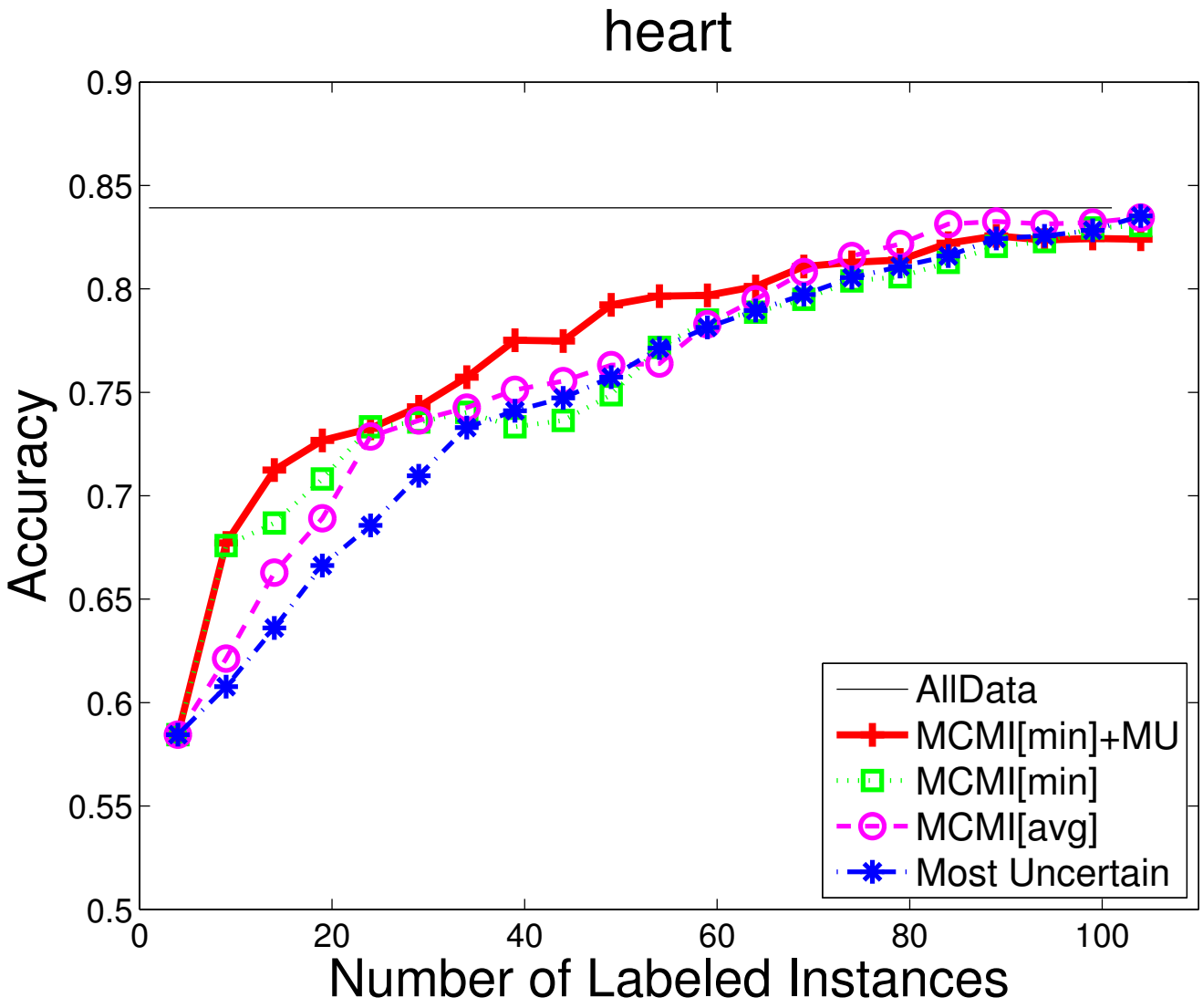

crx
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

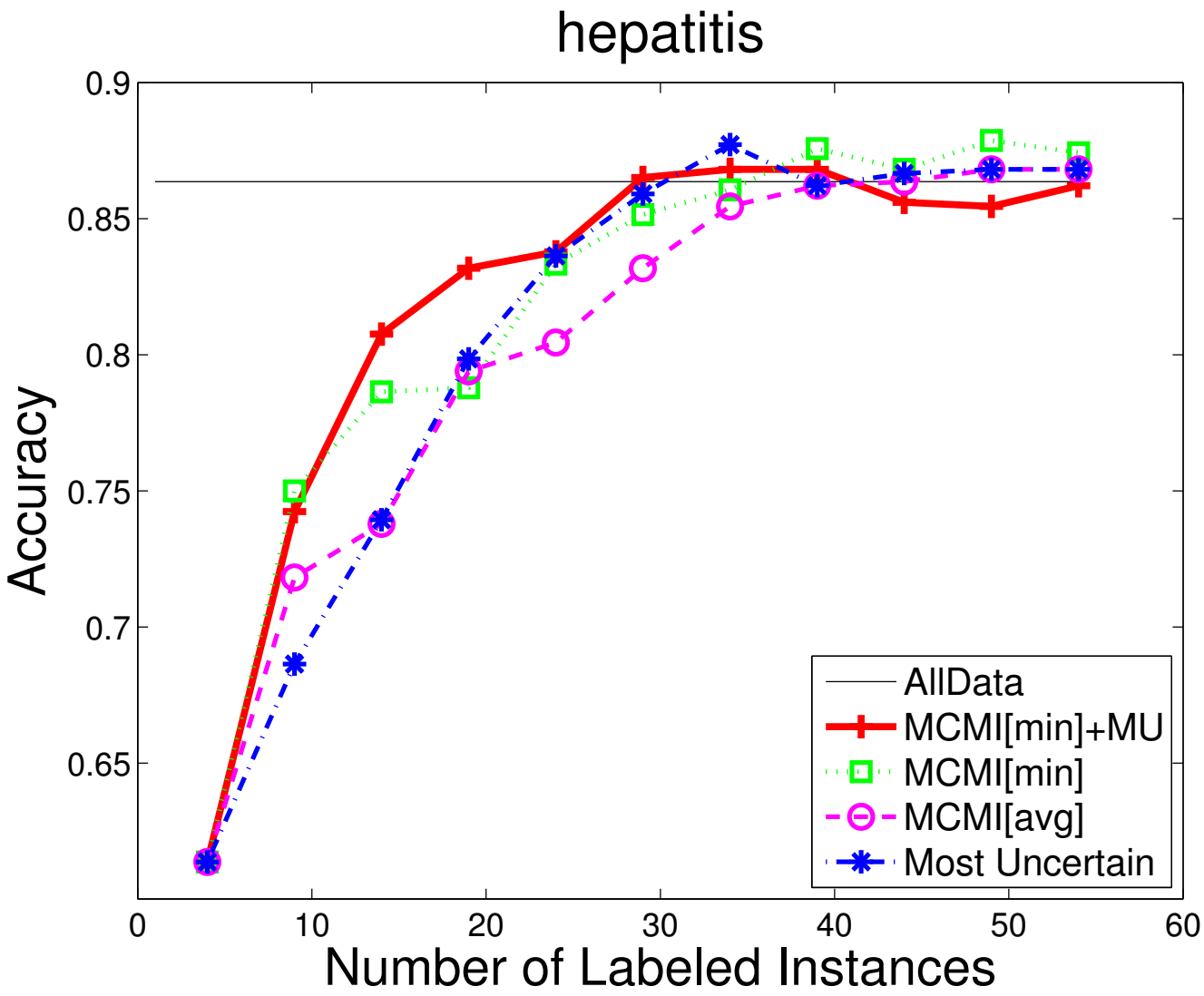

diabetes
Accuracy
Number of Labeled Instances
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

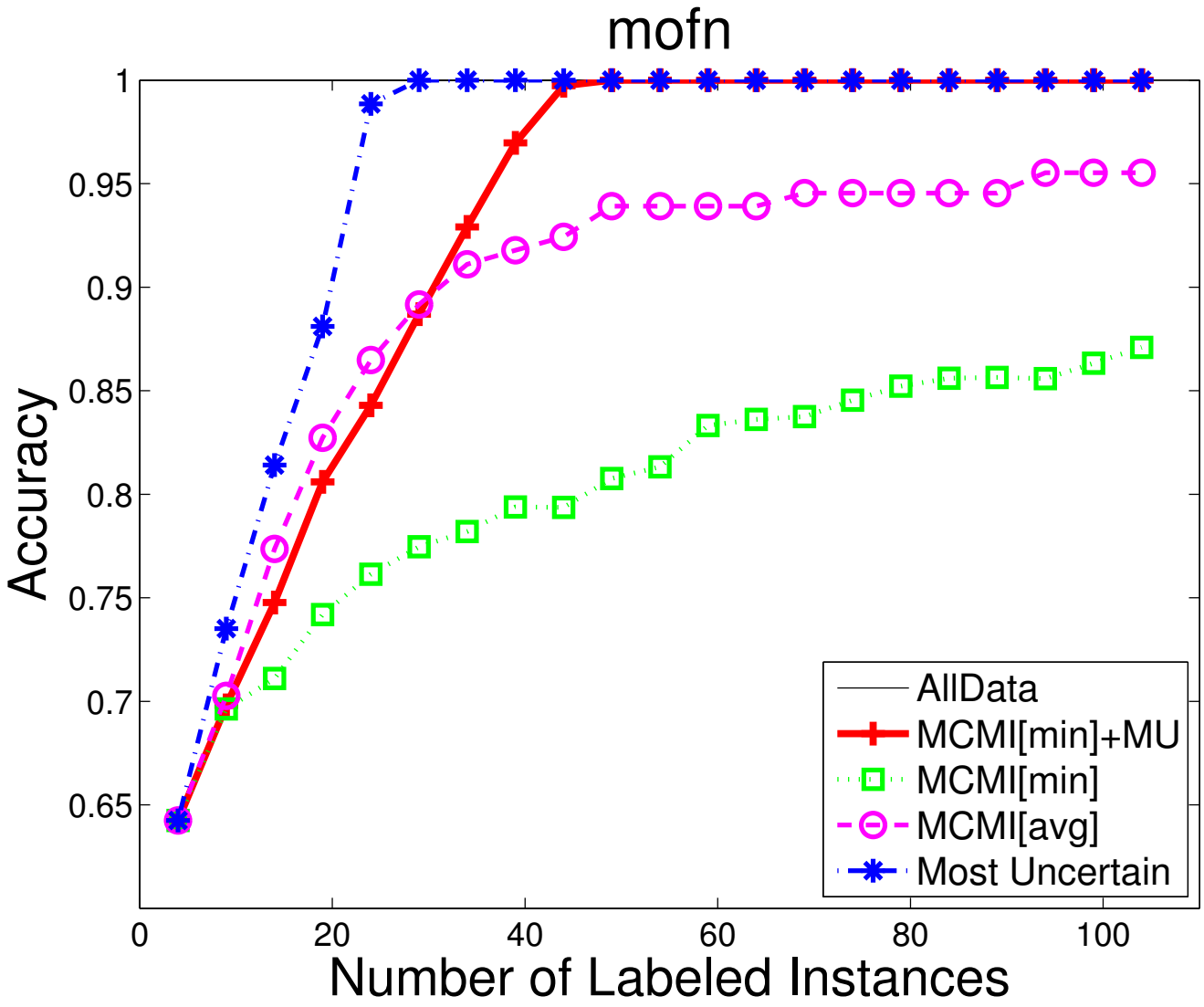

flare
Accuracy
Number of Labeled Instances
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

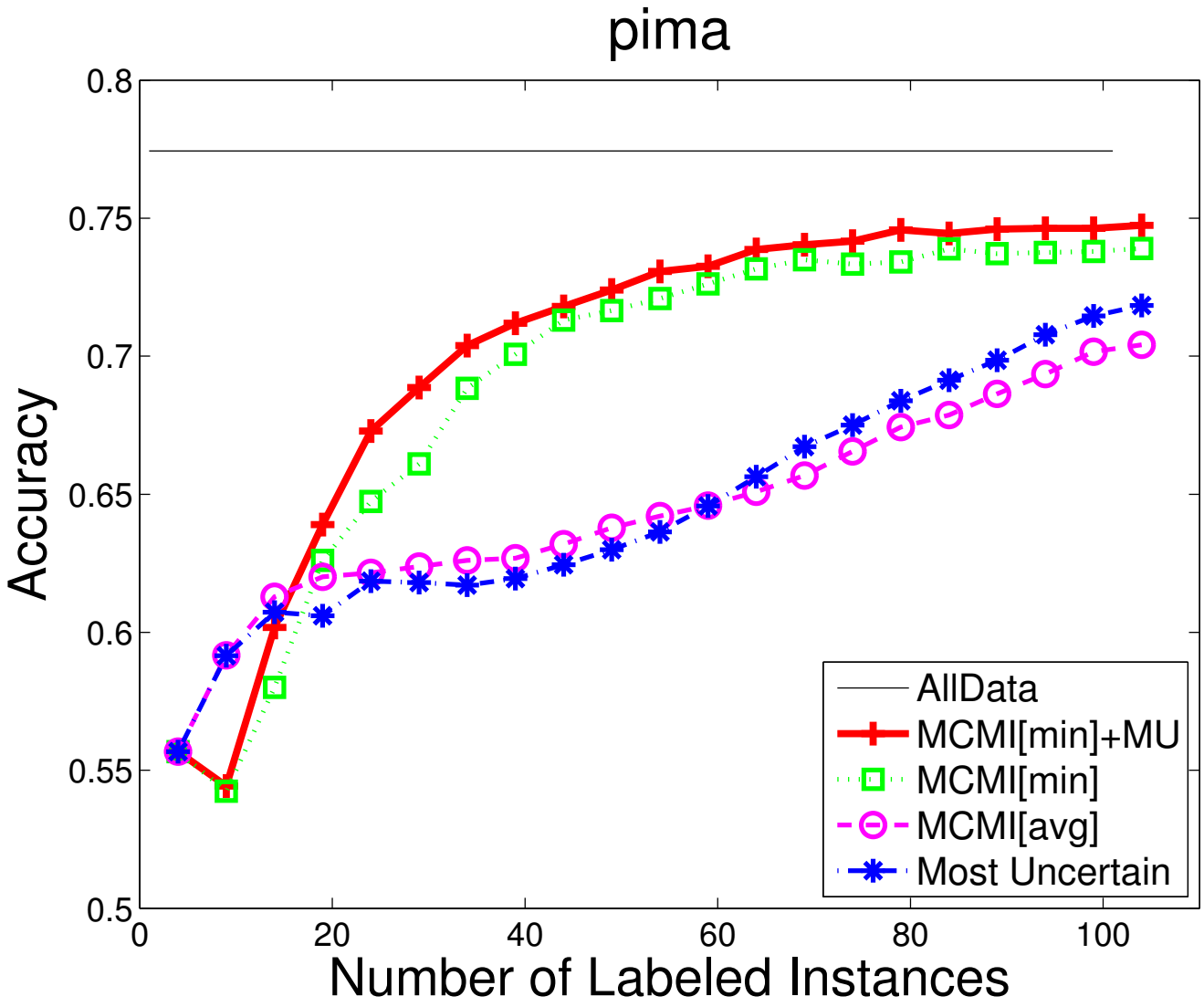

german
Accuracy
Number of Labeled Instances
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

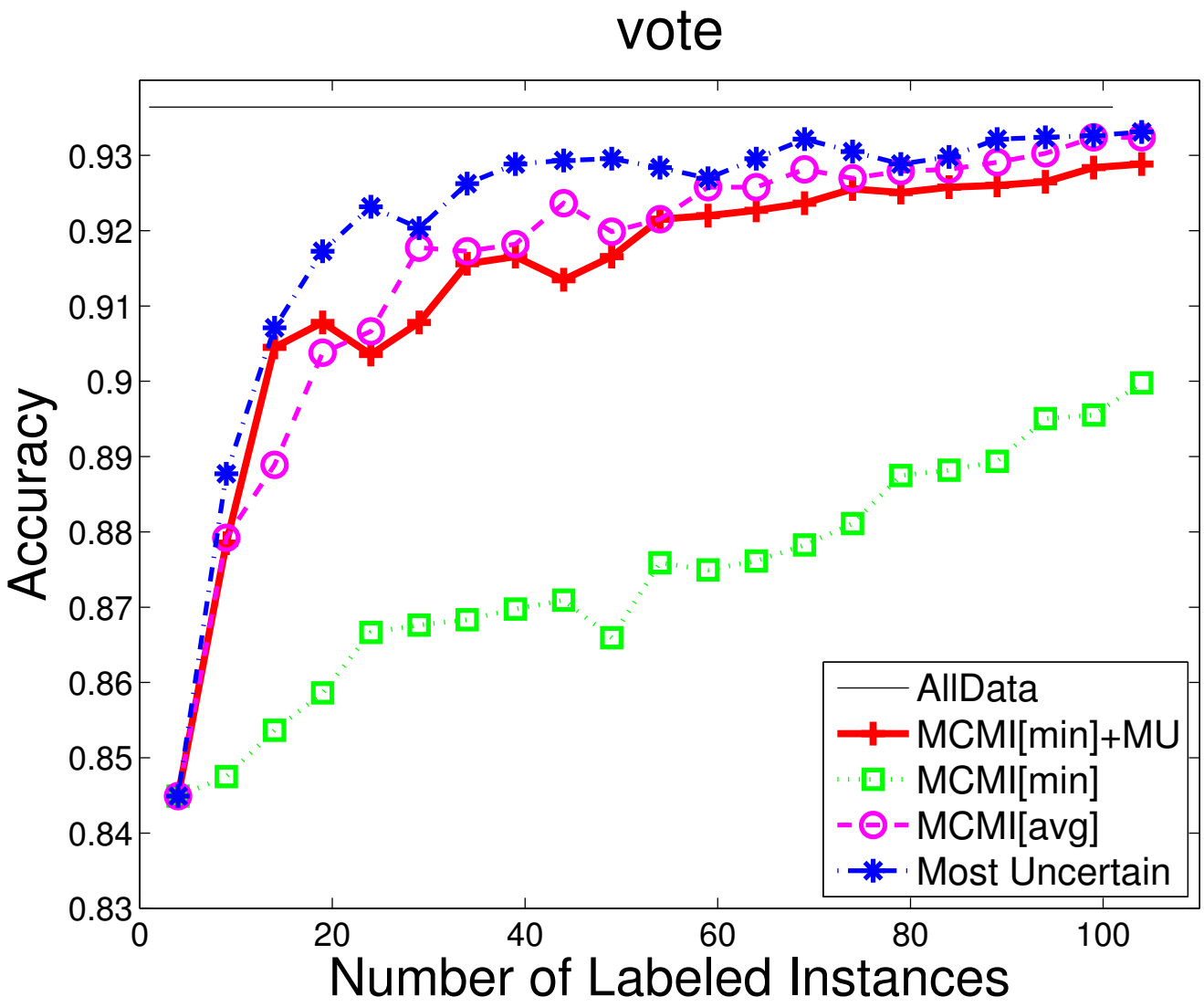

glass2
0.75
0.7
0.65
0.6
0.55
Accuracy
0
20
40
60
80
100
Number of Labeled Instances
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

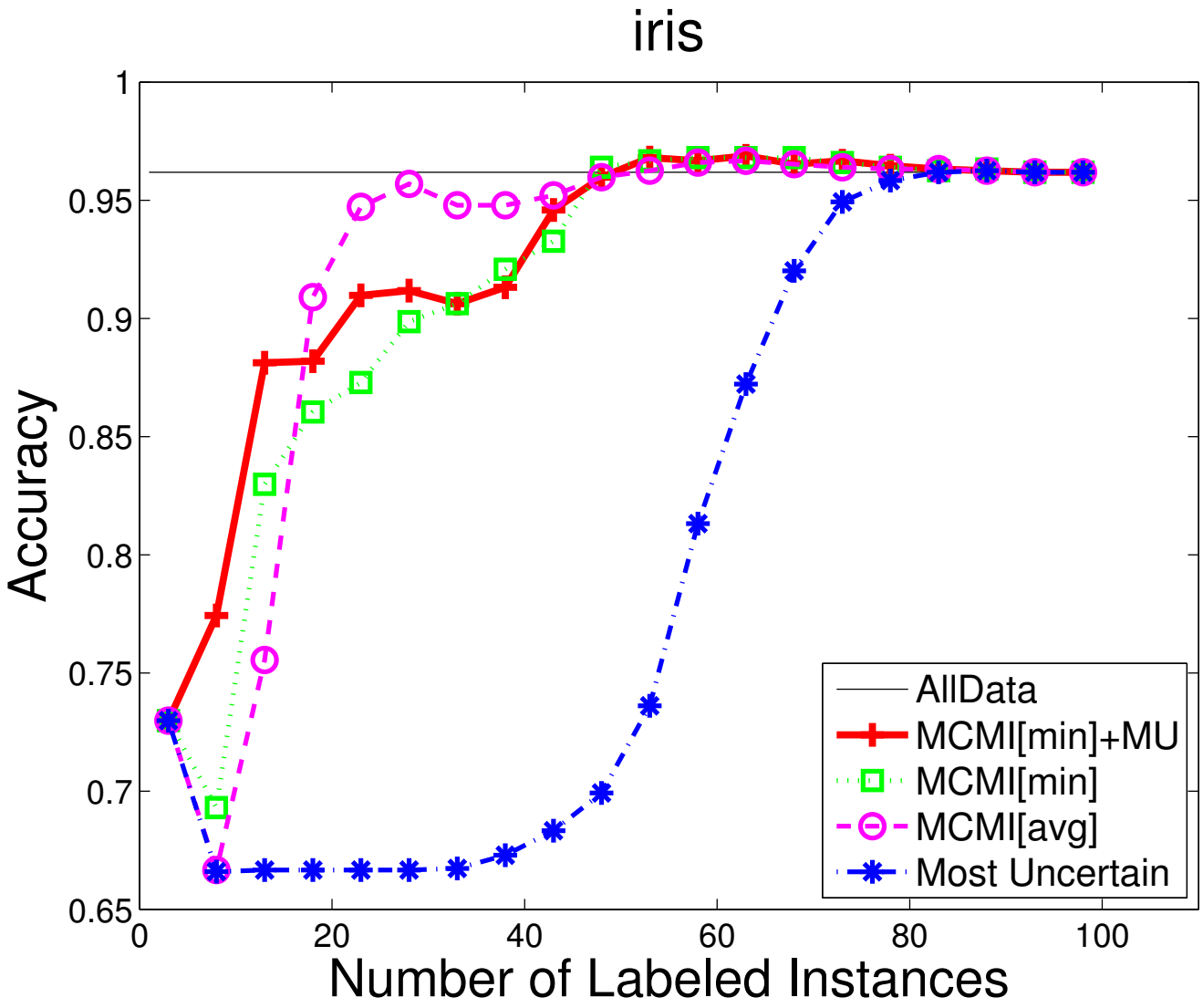

heart
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

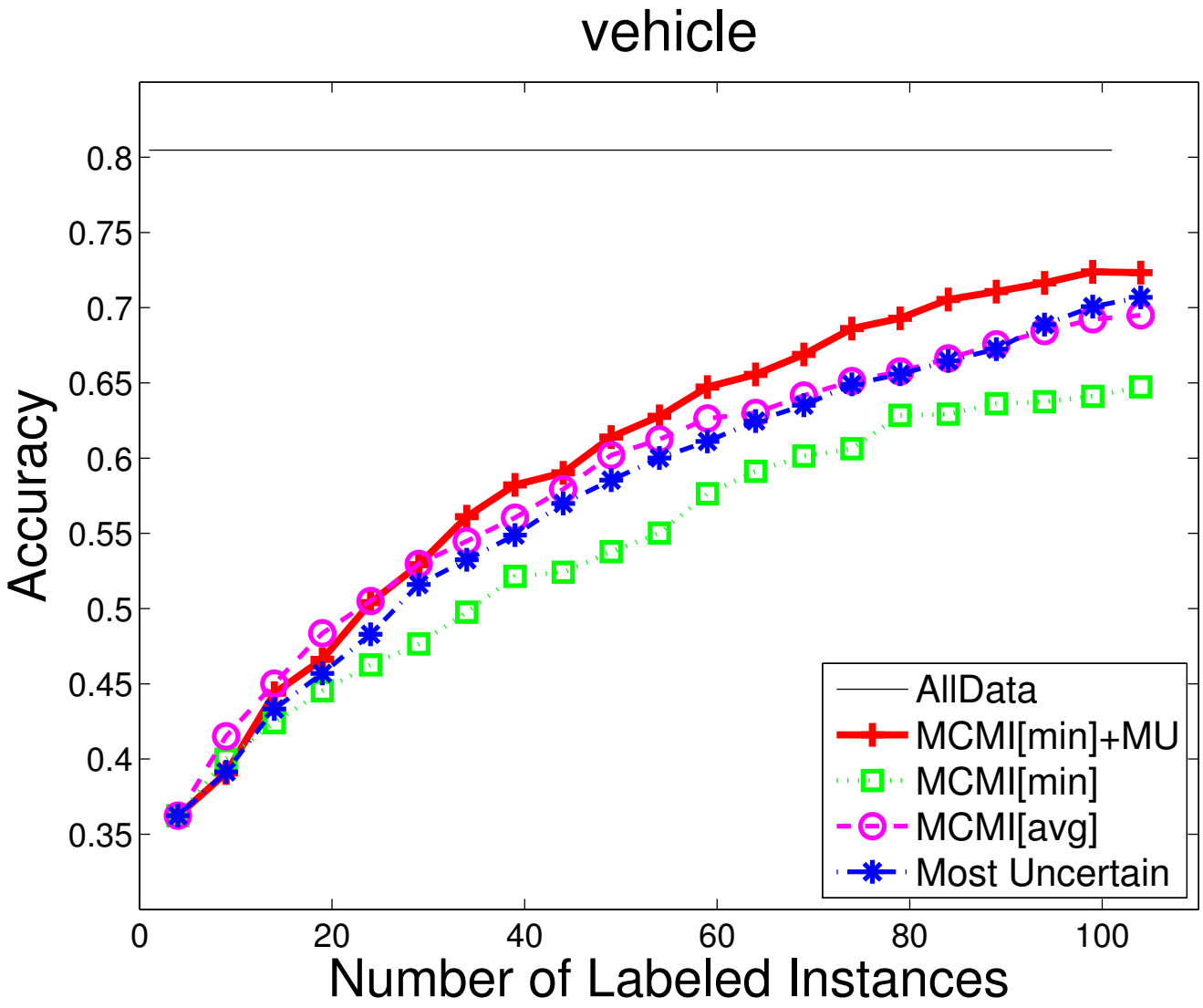

hepatitis
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0
10
20
30
40
50
60
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain

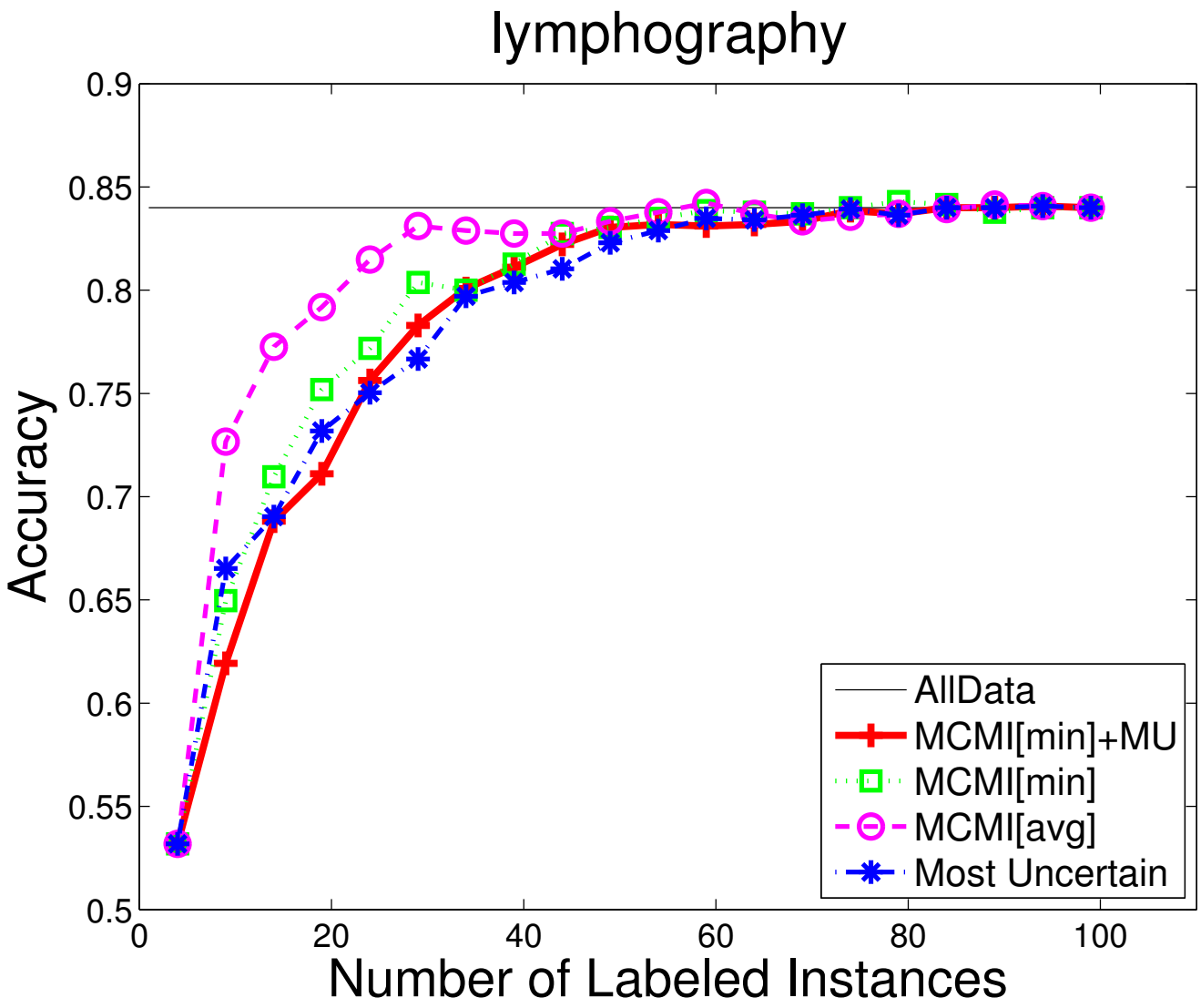

mofn
Accuracy
Number of Labeled Instances
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain


pima
Accuracy
Number of Labeled Instances
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain


vote
Accuracy
Number of Labeled Instances
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain


iris
Accuracy
Number of Labeled Instances
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain


vehicle
Accuracy
Number of Labeled Instances
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain


lymphography
Accuracy
Number of Labeled Instances
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
AllData
MCMI[min]+MU
MCMI[min]
MCMI[avg]
Most Uncertain