The application of Tree-based Methods on the Analysis of MathSciNet Database

↓ classification or regression

A rumor: International collaboration ↑s the P of being cited for an academic work

Aim: to check this hypothesis

Data

Level	Citations	# of samples	# of publications
5	>30	4	54
4	11~30	23	298
3	6~10	23	413
2	2~5	91	1008
1	1	48	628
0	0	116	1506
	Total	305	3907

(works published in 2009 with ≥1 author from JP institute)
Date of late 2018

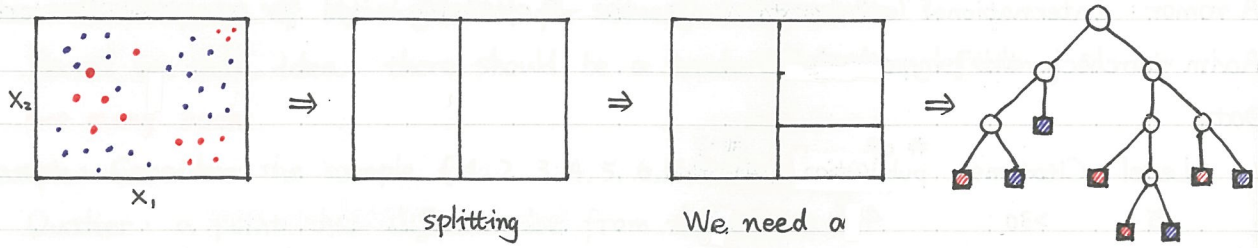
Features considered / Predictors

1.1 # of citation
1.2 level
1.3 # of citation given by ≥1 authors
1.4 # of ... by other authors
3.1 subject
3.2 % of each subject during 2004~8
4.1 # of references given by database
4.2 # of pages
4.3 whether there is a review text
2.1 # of authors
2.2 # of authors publishing 1st work in 2009
2.3 % of young authors
2.4 academac age
2.5 # of clifrent institutes
2.6 # of nations
2.7 % of authors from JP institutes
5.1 Math Citation Quotient for 2009
5.2 Accumulated citations of the journal until 2009

Results

- Works collaborated with international authors (and Japanese authors)
  - are generally more cited than works from only Japanese authors
- Whether # of authors >1 or =1 does not affect the citation
-

Experiment (classification tree)



splitting rule

We need a stoping rule

62 leaves acquired

Results

Predictors	# splits	Total σ^2 drop	σ^2 drop per split
ref	4	89	22
cito2009	5	43	9
mcq	8	54	7
msccla	7	41	6
rfstp	2	11	6
⋮			
Total	61	400	7

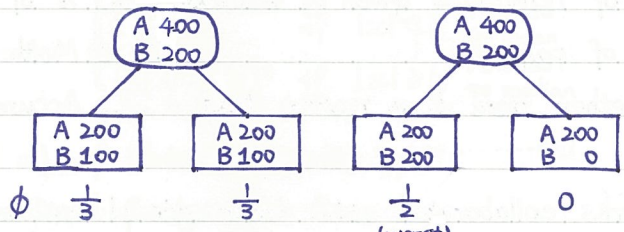
(only 10 predictors used)

Shuffle: put wrong values of one certain predictors
tree → count the ↑ of misclassification rate

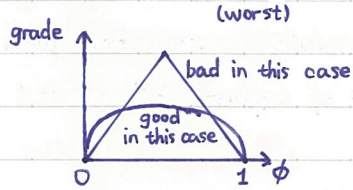
Problem: affected by the position in the tree

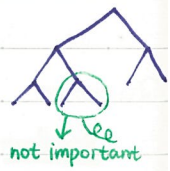
ref	0.21
mcq	0.20
avacag	0.15
pg	0.12
⋮	
nati	0.004

Grading of splitting rules



truth / predict	H lvl	M lvl	L lvl
H lvl	21	2	3
M lvl	4	94	10
L lvl	2	18	151





Problem in cutting the tree from a overfitting complete one