Сравнение эффективности различных методов машинного обучения

Поиск кода с запашком

Код с запашком (code smell)

- Стрельба дробью
- Длинный список параметров
- Теоретическая общность
- Параллельные иерархии наследования

Мотивация

- Есть множество исследований эффективности отдельных методов
- Нет чёткого общепринятого определения запахов кода
- Многие исследования ограничиваются маленькой обучающей выборкой

Масштаб исследования

- В качестве обучающей выборки использовано 74 различный java-проекта
- 32 различных алгоритма
- 4 запаха кода
- Полный перебор дополнительных параметров алгоритмов

Запахи кода

Запахи уровня классов:

- Большой класс (*God/Large/ Brain Class*)
- Класс данных (*Data class*)

Запахи уровня методов:

- Завистливая функция (*Feature Envy*)
- Длинный метод (*Long/God Method*)

Метрики

Size	Complexity	Cohesion	Coupling	Encapsulation	Inheritance
LOC	CYCLO	LCOM5	FANOUT	LAA	DIT
LOCNAMM*	WMC	TCC	ATFD	NOAM	NOI
NOM	WMCNAMM*		FDP	NOPA	NOC
NOPK	AMWNAMM*		RFC		NMO
NOCS	AMW		CBO		NIM
NOMNAMM*	MAXNESTING		CFNAMM*		NOII
NOA	WOC		CINT		
	CLNAMM		CDISP		
	NOP		MaMCL§		
	NOAV		MeMCL§		
	ATLD*		NMCS§		
	NOLV		CC		
			CM		

Метрики (2)

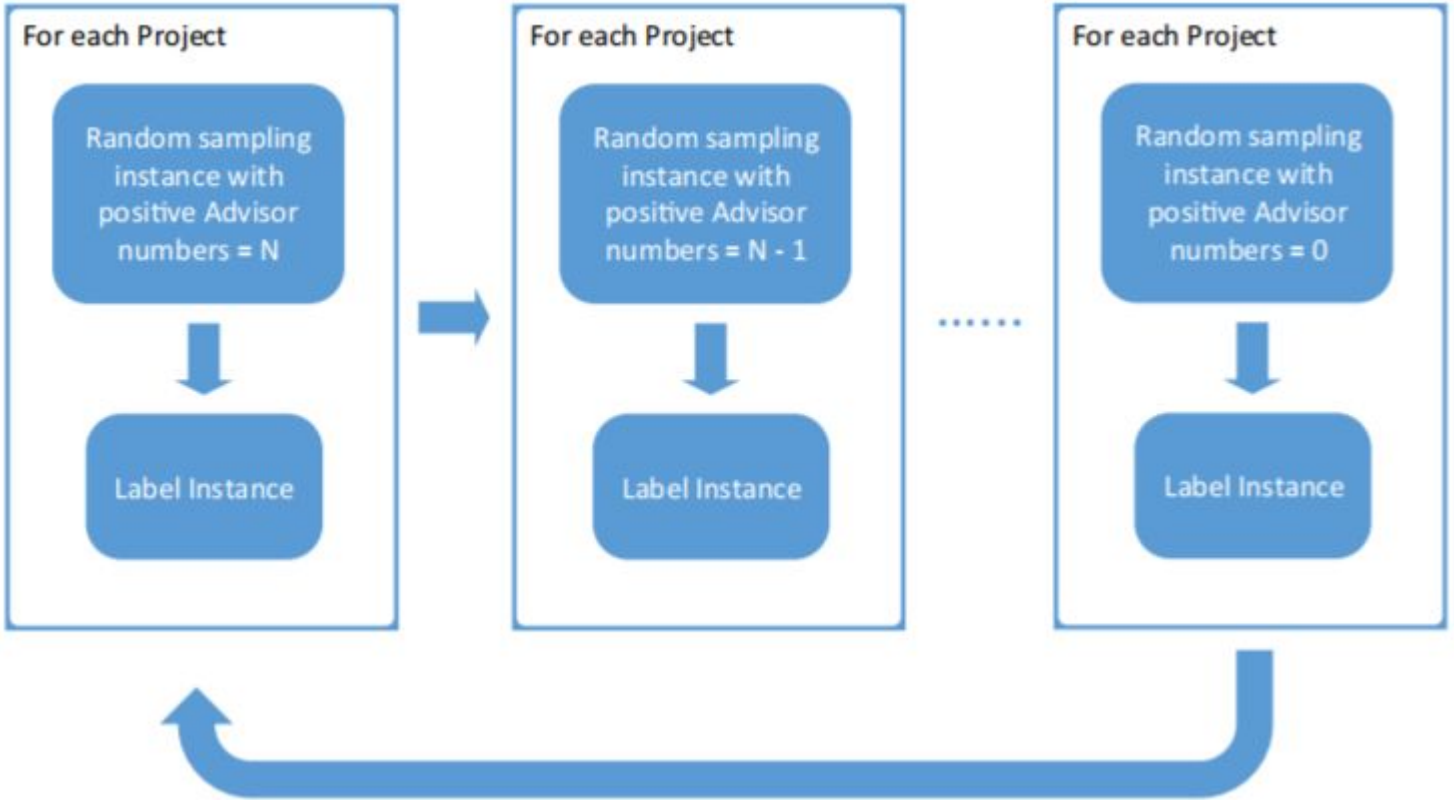
NODA	NOPVA	NOPRA	NOFA
NOFSA	NOFNSA	NONFNSA	NOSA
NONFSA	NOABM	NOCM	NONCM
NOFM	NOFNSM	NOFSM	NONFMABM
NONFNSM	NONFSM	NONAM	NOSM
NOPLM	NOPRM	NOPM	NODM

Обучающая выборка

- Большой объём данных
- Несбалансированность данных

Code smell	Advisors: detection tools or rules
God Class	iPlasma, PMD
Data Class	iPlasma, Fluid Tool, Antipattern Scanner
Long Method	iPlasma, PMD, Marinescu (2002)
Feature Envy	iPlasma, Fluid Tool

Обучающая выборка



Обучающая выборка

Класс данных

- не должен содержать сложных методов
- основную массу методов составляют геттеры
- может содержать несколько не-геттеров, но достаточно простых
- каждый атрибут должен быть либо публичным либо быть доступным через геттер

Длинный метод

- должен быть сложным
- должен быть длинным
- должен иметь много аргументов
- должен использовать большое число атрибутов своего класса, а также локальных переменных и атрибутов, полученных через геттеры

Обучающая выборка

Number of advisors indicating a smell	Number / percentage of smelly instances on datasets			
	Data class	God class	Feature envy	Long method
3	2 / 0.004 %	102 / 0.199 %	0	166 / 0.044 %
2	286 / 0.587 %	425 / 0.829 %	587 / 0.157 %	1514 / 0.404 %
1	4377 / 8.987 %	1319 / 2.573 %	89,580 / 23.902 %	3693 / 0.985 %
0	44,041 / 90.422 %	49,418 / 96.399 %	284,621 / 74.942 %	369,415 / 98.566 %

Обучающая выборка

Number of advisors indicating a smell	Number of SMELLY/NON-SMELLY instances on datasets			
	Data class	God class	Feature envy	Long method
3	2/0	69/1	-	84/0
2	97/22	109/18	136/8	133/15
1	36/112	22/126	56/169	32/116
0	9/152	4/144	37/111	0/166

Представленные алгоритмы

- J48 Pruned
- J48 Unpruned
- J48 Reduced Error Pruning
- JRip
- Random Forest
- Naïve Bayes
- SMO RBF Kernel
- SMO Poly Kernel

- LibSVM C-SVC Linear Kernel
- LibSVM C-SVC Polynomial Kernel
- LibSVM C-SVC Radial Kernel
- LibSVM C-SVC Sigmoid Kernel
- LibSVM ν-SVC Linear Kernel
- LibSVM ν-SVC Polynomial Kernel,
- LibSVM ν-SVC Radial Kernel
- LibSVM ν-SVC Sigmoid Kernel

Подбор параметров

- Реализация алгоритма в weka - чёрная коробка
- Поиск по сетке
  - Легко параллелится
  - Более-менее полное покрытие пространства возможных комбинаций
- 10-блочная кросс-валидация, повторённая 10 раз
- Для SVM - варианты предобработки данных
- Параметр сравнения - точность
- Попарное сравнение (победа/поражение/ничья)

Результаты (класс данных)

Classifier	Wilcoxon victories		Cliff Delta effect size	
	Number	Percentage	Value	Magnitude
B-J48 Pruned	29	94 %	0,48	large
Random Forest	26	84 %	0,49	large
B-JRip	26	84 %	0,43	medium
B-J48 Unpruned	24	77 %	0,40	medium
B-Random Forest	24	77 %	0,37	medium
J48 Pruned	23	74 %	0,37	medium
J48 Unpruned	20	65 %	0,32	small
JRip	17	55 %	0,28	small
B-J48 Reduced Error Pruning	16	52 %	0,27	small
J48 Reduced Error Pruning	14	45 %	0,25	small

Результаты (Большой класс)

Classifier	Wilcoxon victories		Cliff delta effect size	
	Number	Percentage	Value	Magnitude
Naïve Bayes	24	77 %	0,33	medium
J48 Pruned	21	68 %	0,29	small
J48 Unpruned	21	68 %	0,29	small
J48 Reduced Error Pruning	21	68 %	0,28	small
Random Forest	20	65 %	0,29	small
B-J48 Pruned	20	65 %	0,29	small
B-J48 Reduced Error Pruning	20	65 %	0,28	small
JRip	20	65 %	0,25	small
B-Random Forest	19	61 %	0,22	small
B-Naïve Bayes	19	61 %	0,21	small

Результаты (завистливая функция)

Classifier	Wilcoxon victories		Cliff delta effect size	
	Number	Percentage	Value	Magnitude
B-J48 Pruned	26	84 %	0,46	medium
B-J48 Unpruned	25	81 %	0,43	medium
B-JRip	23	74 %	0,44	medium
B-Random Forest	23	74 %	0,42	medium
J48 Unpruned	23	74 %	0,37	medium
Random Forest	22	71 %	0,39	medium
J48 Reduced Error Pruning	20	65 %	0,35	medium
B-J48 Reduced Error Pruning	19	61 %	0,34	medium
J48 Pruned	18	58 %	0,35	medium
SMO Poly Kernel	16	52 %	0,28	small

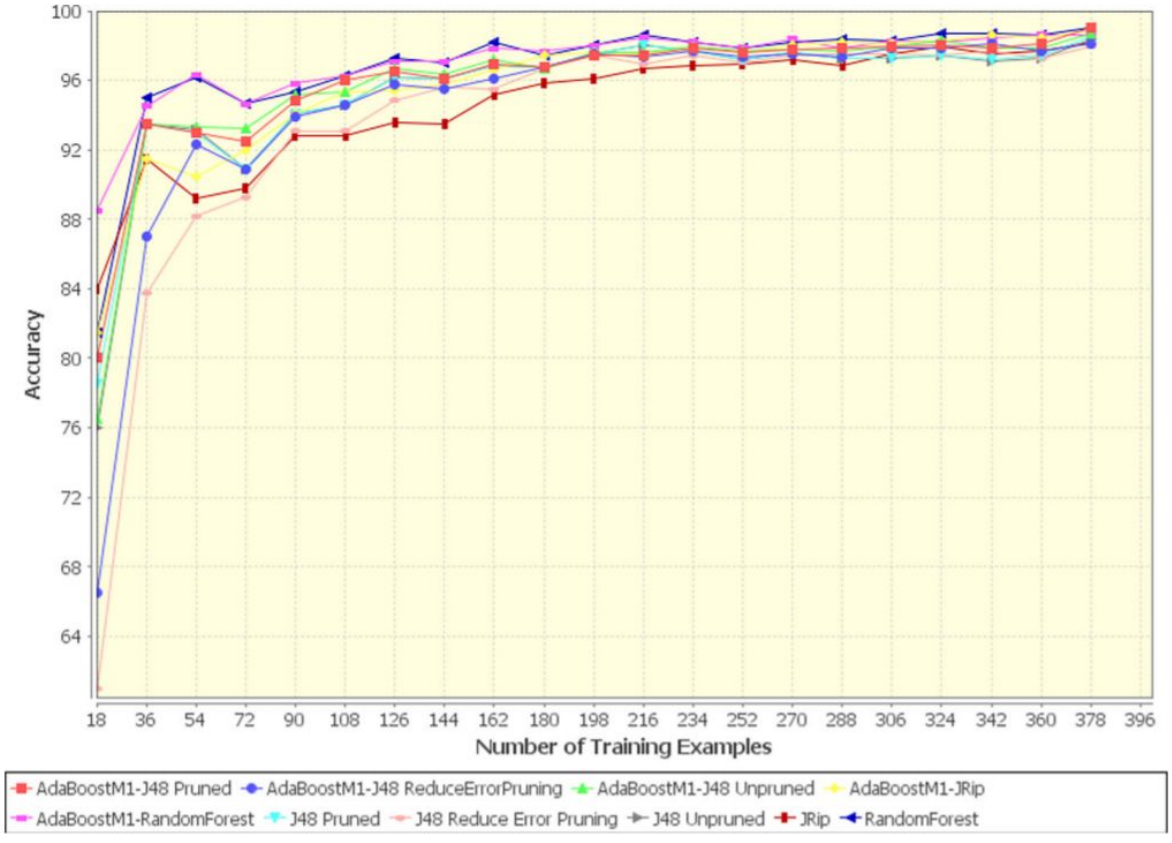
Результаты (длинный метод)

Classifier	Wilcoxon victories		Cliff delta effect size	
	Number	Percentage	Value	Magnitude
B-J48 Unpruned	25	81 %	0,46	medium
B-Random Forest	25	81 %	0,43	medium
B-J48 Pruned	25	81 %	0,40	medium
B-JRip	25	81 %	0,40	medium
Random Forest	24	77 %	0,42	medium
B-J48 Reduced Error Pruning	23	74 %	0,36	medium
JRip	22	71 %	0,32	small
J48 Pruned	18	58 %	0,33	small
J48 Unpruned	16	16 %	0,32	small
SMO Poly Kernel	13	42 %	0,21	small

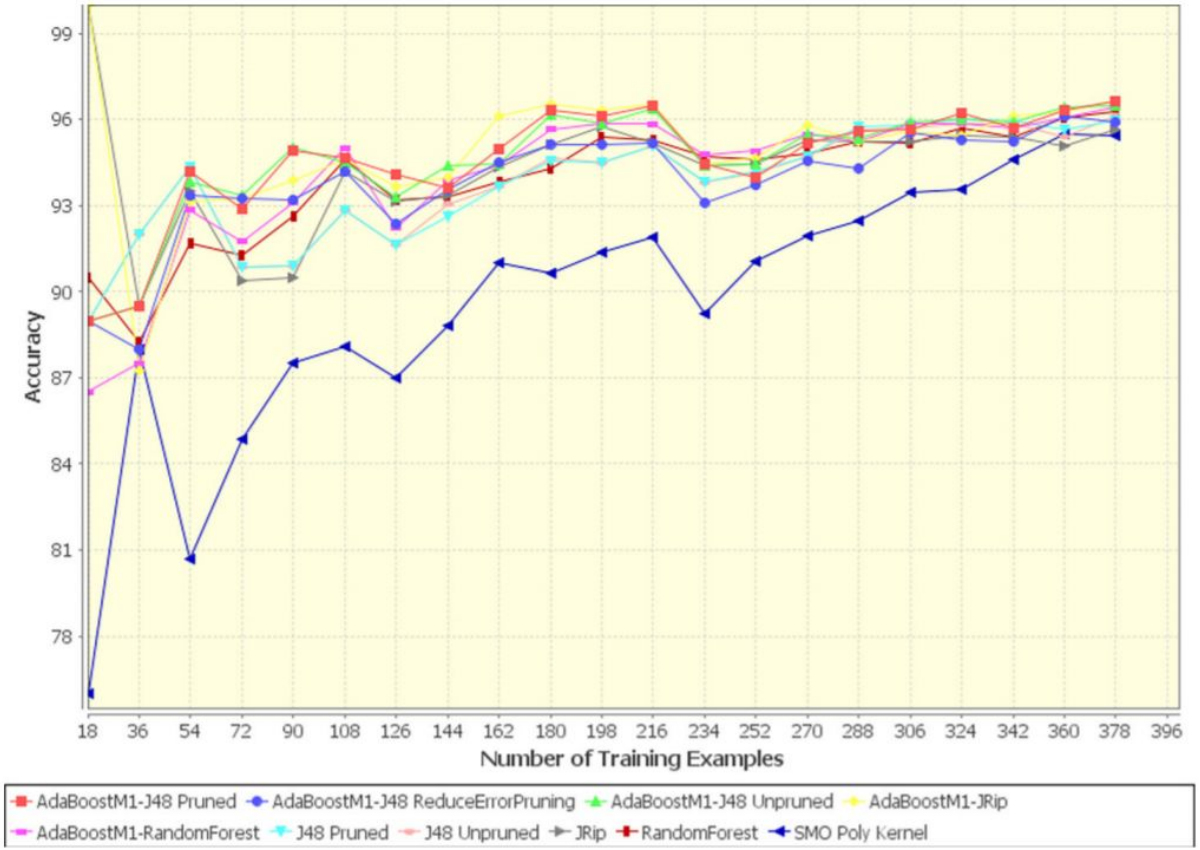
Результаты

Classifier	Data class	God class	Feature envy	Long method
B-J48 Pruned	1	6	1	3
B-J48 Reduced Error Pruning	9	7	8	6
B-J48 Unpruned	4	–	2	1
B-JRip	3	–	3	4
B-Naïve Bayes	–	10	–	–
B-Random Forest	5	9	4	2
J48 Pruned	6	2	9	8
J48 Reduced Error Pruning	10	4	7	-
J48 Unpruned	7	3	5	9
JRip	8	8	–	7
Naïve Bayes	–	1	–	–
Random Forest	2	5	6	5
SMO Poly Kernel	–	–	10	10

Зависимость от размера выборки



Зависимость от размера выборки



Пример результатов обучения

Code smell	Rules By J48 pruned	Rules by JRip	Comment
Data class	NOAM>2 and WMCN AMM≤21 and NIM≤30	(WOC<0.352941 and NOAM≥4 and RFC≤41) or (CFNAMM=0 And NOAM≥3) or (AMW≤1 And NOPVA≥3)	NIM and RFC are not conceptually part of Data Class smell.
God class	WMCNAMM≥48	WMCNAMM≥48	Both algorithms produce an equal rule
Feature Envy	ATFD(method)>4 and LAA<0.458571 and NOA≤16	(ATFD≥5 And LAA<0.3125) Or ATFD≥9 Or (FDP≤3 And NMO≤1)	NOA and NMO are not conceptually part of Feature Envy smell.
Long Method	LOC(method)≥80 And CYCLO≥10	LOC method≥80 And CYCLO≥8	Both algorithms produce a nearly equal rule

Результаты запусков на всей выборке

Code smell	Winner algorithm	Highest % algorithm	Lowest % algorithm
Data Class	B-J48 Pruned (4.58 %)	B-LibSVM C-SVC Sigmoid Kernel (12.33 %)	B-JRip (3.83 %)
God Class	Naïve Bayes (9.63 %)	B-LibSVM C-SVC Polynomial Kernel (12.37 %)	LibSVM C-SVC Sigmoid Kernel (4.29 %)
Feature Envy	B-JRip (3.37 %)	B-SMO RBF Kernel (30.24 %)	J48 Pruned (3.01 %)
Long Method	B-J48 Pruned (1.21 %)	B-LibSVM C-SVC Polynomial Kernel (20.47 %)	B-J48 Pruned (1.21 %)