Статистическое сравнение классификаторов

А. Катруца

МФТИ

24 марта 2014 г.

Постановка задачи

Есть много классификаторов, много выборок и функционал качества. Хотим понять какие классификаторы отличаются по эффективности.
Предположения:

- результатам измерений можно верить
- оценка качества для каждого алгоритма была получена на одинаковых случайных взятых подвыборках
- схема сэмплирования неизвестна.

Используемые функционалы качества

	1999	2000	2001	2002	2003
Total number of papers	54	152	80	87	118
Relevant papers for our study	19	45	25	31	54
Sampling method [%]					
cross validation, leave-one-out	22	49	44	42	56
random resampling	11	29	44	32	54
separate subset	5	11	0	13	9
Score function [%]					
classification accuracy	74	67	84	84	70
classification accuracy - *exclusively*	68	60	80	58	67
recall, precision...	21	18	16	25	19
ROC, AUC	0	4	4	13	9
deviations, confidence intervals	32	42	48	42	19
Overall comparison of classifiers [%]	53	44	44	26	45
averages over the data sets	0	4	6	0	10
t-test to compare two algorithms	16	11	4	6	7
pairwise t-test one vs. others	5	11	16	3	7
pairwise t-test each vs. each	16	13	4	6	4
counts of wins/ties/losses	5	4	0	6	9
counts of *significant* wins/ties/losses	16	4	8	16	6

Усреднение по выборкам

Недостатки:

- чувствительность к выбросам
- усреднение несравнимых результатов бессмысленно

Считается неэффективным методом сравнения.

Парный t-test

Пусть c_i^1 и c_i^2 качества алгоритмов a_1 и a_2 на i-ой выборке и $d_i = c_i^1 - c_i^2$. N — количество выборок.

$$H_0 : a_1 \overset{c}{\sim} a_2$$
$$H_1 : \text{not } H_0$$

$\overline{d}/\sigma_{\overline{d}}^2 \sim \chi^2(N-1)$ при справедливости H_0.
В R функция: t.test(y1,y2,paired=TRUE).
Проблемы:

- Соразмерность получаемых результатов.
- Предположение нормальности разницы средних, трудно проверить.
- Чувствителен к выбросам.

Критерий знаковых рангов Уилкоксона

В R: wilcox.test(y1, y2, paired=TRUE)
Достоинства:

- требует только качественную соразмерность (абсолютный разброс значений игнорируется)
- не предполагает нормальность
- менее чувствителен к выбросам

Используется при нарушении предположений t-test'а.

Количество выигрышей и проигрышей

- Не предполагается нормальность или соразмерность, однако намного слабее предыдущего критерия.
- Нельзя отбрасывать какие-то измерения, ссылаясь на их случайный характер, тест не умеет определять, где случайный, а где значимый.
- Отбор только значимых выигрышей или проигрышей ведёт к ослаблению критерия.

ANOVA

Предположения:

1. Выборки из нормального распределения
2. Сферичность — равенство дисперсий

При отвержении H_0 используются post-hoc tests:

1. Tukey test для сравнения всех со всеми
2. Dunnet test для сравнения всех с эталоном.

Tukey: статистика — аналогична t-test, но берётся максимальный разброс средних.
В R: TukeyHSD(fit).
Dunnet test: статистика — наибольшая по модулю из t-статистик для каждого алгоритма.
В R: dunnett.test(Z = Z, select = rep(1, length(Z))), пакет asd.

Критерий Фридмана

r_i^j — ранг j-го алгоритма на i-ой выборке. $R_j = \frac{1}{N}\sum_i r_i^j$

Если H_0 верна, R_j близки.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \sim F(k-1,(k-1)(N-1))$$

Используется при нарушении предположений ANOVA.

В R: friedman.test(x)

При отклонении нулевой гипотезы используются post-hoc tests.

Критерий Неменьи

Сравниваем всех со всеми, аналогично Tukey test.
Алгоритмы считаются различными, если средняя разность рангов превышает критическую разность:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}},$$

где q_α — стьюдентизированная порядковая статистика, поделённая на $\sqrt{2}$ — см. в таблицах.

Критерий Бонферрони-Данна

В случае сравнения с эталоном используются варианты критериев из множественной проверки гипотез.
Статистики те же, что и в критерии Неменьи, но критическая разность считается для $\frac{\alpha}{k-1}$.
Post-hoc test маломощные, критерий Фридмана может показать существенное различие, а они нет.

Пример

Индексы m и cf показывают настройку параметров: минимального числа элементов в листе и доверительного интервала. Значения по умолчанию: m = 0, cf = 0.25.

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

Пример

Применим критерий Фридмана:

$$\chi_F = 9.28 \qquad F_F = 3.69$$

Критическое значение для $F(3, 39)$ при $\alpha = 0.05$ равно 2.85. Следовательно, H_0 отклоняем.
Дальнейшее применение критерия Неменьи показывает, что настройка m и m+cf улучшает качество алгоритма, а об эффективности настройки cf никакого вывода сделать нельзя.