Принципы и методы эксплоративного выделения и анализа подгрупп пациентов для персонализации лечения.

Наталья Корепанова
Научный руководитель: д.ф.-м.н., проф. С.О. Кузнецов

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

20 октября 2017 г.

Основные обозначения

$\mathcal{X}$ - пространство значений инициальных признаков пациентов размерности $p \in \mathbb{N}$
$\mathcal{T} = \{0, ..., q\}$ - доступные стратегии лечения, $q \in \mathbb{N}$
$\mathcal{Y}$ - множество допустимых ответоов на лечение (бинарный или количественный)
$f(X, T) = E(Y|X, T)$ - функция ответа пациента на лечение, где X, T, Y - случайные величины, принимающие значение из $\mathcal{X}, \mathcal{T}$ и $\mathcal{Y}$ соответственно.

Известно n реализаций троек (x_i, t_i, y_i), где $x_i \in \mathcal{X}$, $t_i \in \mathcal{T}$, $y_i \in \mathcal{Y}$, $i = 1, ..., n$.

Общий вид задачи персонализации

По имеющимся наблюдениям для любого $x \in \mathcal{X}$ выбрать $t \in \mathcal{T}$ такое, что $f(x,t) = \max_{t' \in \mathcal{T}} f(x,t')$.

Задача выделения подгрупп пациентов для персонализации бинарного лечения

Пусть $q = 1$, т.е. $\mathcal{T} = \{0,1\}$, а X_i и T_i независимы, $i = 1,\ldots,n$. Найти такие области (подгруппы пациентов) $\mathcal{X}'$ пространства $\mathcal{X}$, в которых $|f(X,0) - f(X,1)|$ "значимо больше" 0. Такие подгруппы будем называть хорошими.

Проблема

Рост вероятности ошибиться, назвав подгруппу хорошей, с ростом числа подгрупп-кандидатов.

Например, если вероятность такой ошибки при проверке одной подгруппы составляет 0.05, то при проверке m независимых таких подгрупп она составит $1 - (1 - 0.05)^m$
(при $m = 5$ она примерно равна 0.2, при $m = 10$ – 0.4).

Первые попытки осмысления

25 правил (Brookes ST et al., 2001), 21 правило (Rothwell PM, 2005), 11 критериев оценки надёжности результатов (Sun X et al., 2010)

Основные идеи

1. Подгруппы задаются до начала сбора данных (проведения рандомизированного исследования).
2. Подгруппы должны быть биологически правдоподобны.
3. Поправка на множественность сравнений.
4. Нельзя проводить анализ подгрупп, если различия в ответе на разное лечение статистически значимы на всей выборке.
5. К результатам анализа подгрупп стоит относиться с осторожностью.

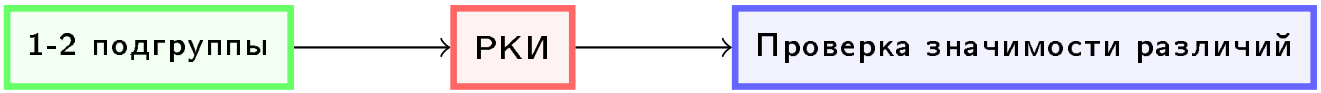
Food and Drug Administration (2012), European Medicines Agency (2014).

Анализ подгрупп

(Mayer C et al., 2015)

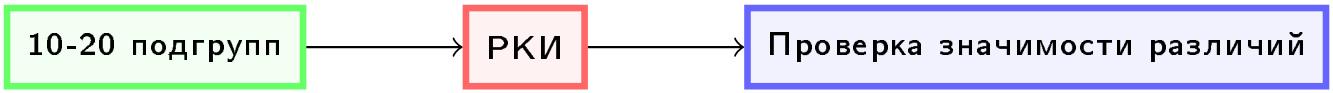
РКИ - рандомизированное контролируемое исследование

Конформативный анализ подгрупп

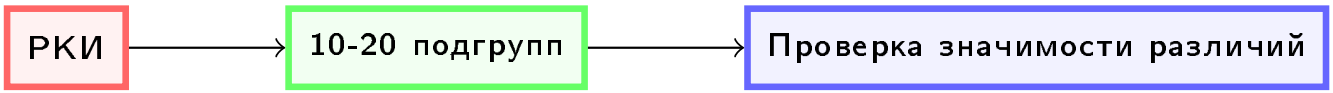


Эксплоративный анализ подгрупп

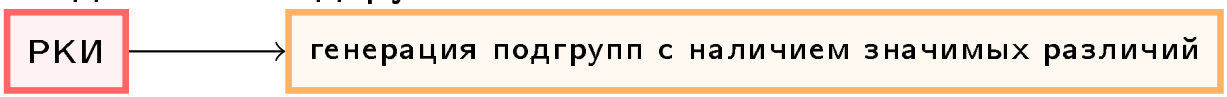
Эксплоративная оценка подгрупп



Post-hoc оценка подгрупп



Выделение подгрупп



Выделение подгрупп

Исследование и выделение подгрупп = Отбор моделей

Методы отбора с поправками на множественность сравнений

1. Из большого пространства моделей выделения подгрупп выбирается модель, выделяющая подгруппы со значимыми различиями.
2. Мета-параметры оцениваются по данным.
3. Пространство допустимых моделей и методы оценки мета-параметров фиксируются заранее.

Data-driven принципы

(Lipkovich I et al., 2016)

- Контроль переобучения (прунинг и снижение размерности итоговой модели).
- Независимость вероятности ошибочного выделения подгруппы от числа уникальных значений каждого из инициальных признаков пациента.
- Оценка воспроизводимости подгрупп на новых данных (обучающая и тестовая выборки, кросс-валидация).
- Ресемплирование и байесовские методы для получения «честных» оценок различия в ответе на лечение подгруппах.

Таксономия методов

Глобальное моделирование ответа

Восстановление функции $f(X, T)$ на всем пространстве $\mathcal{X} \times \mathcal{T}$.
Различие в ответе при $X = x$ оценивается как $f(x, 1) - f(x, 0)$.

Глобальное моделирование различий

Предположим, что $f(X, T) = h(X, z(X)T)$, где z - монотонная функция, h - монотонная по второму аргументу функция, а $T \in \{0, 1\}$.
Тогда при $X = x$ различия в ответе на терапию между разными лечениями полностью определяется $z(x)$ (Xu Y et al., 2015).
При этом $z(X) = g(f(X, 1), f(X, 0))$.
Восстановление функции $z(X)$ на всем пространстве $\mathcal{X}$.

Локальное моделирование

Нет цели построить модель для всего пространства $\mathcal{X}$.

Глобальное моделирование ответа

Регрессионные модели с регуляризацией

Например, часто предполагается, что $f(X, T) = v(w(X) + l(z(X)T))$, где v и l монотонные функции. Тогда в работе (Imai K et al., 2013) предлагается отдельная регуляция для параметров $w(X)$ и для параметров $l(z(X)T)$.

Методы оценки потенциального ответа

Потенциальный ответ $\tilde{Y}_i(t) = Y_i$, если $T_i = t$, и неизвестен, если $T_i \neq t$.
Оценивается потенциальный ответ (например, при помощи случайного леса (Foster JC et al., 2011)), а затем моделируется функция $f(X, T)$.

Байесовские методы

Глобальное моделирование различий

Методы рекурсивного деления пространства

Функция $z(X)$ моделируется как кусочно постоянная функция. Пространство $\mathcal{X}$ разбивается на подгруппы, покрывающее все пространство $\mathcal{X}$ без попарных пересечений.

Методы моделирования оптимального режима лечения

Режим лечения $d : \mathcal{X} \to \mathcal{T}$. При $\mathcal{T} = \{0, 1\}$ оптимальный режим лечения
$d_{opt}(X) = \max_{d(X)} E(\tilde{Y}(1)d(X) + \tilde{Y}(0)(1 - d(X)))$.

Моделирование аплифта (Uplift modeling)

Аплифт: $m(X) = f(X, 1) - f(X, 0)$.
Методы на основе деревьев, регрессионные методы, SVM методы, ансамблевые методы.

Байесовские методы

Локальное моделирование

На основе bump-hunting

В пространстве $\mathcal{X}$ выделяется область $\mathcal{X}'$ (обычно прямоугольная), такая, что оценка

$$\Big|\big(E(Y|X \in \mathcal{X}', T=1) - E(Y|X \in \mathcal{X}', T=0)\big) - \big(E(Y|X \in \mathcal{X} \setminus \mathcal{X}', T=1) - E(Y|X \in \mathcal{X} \setminus \mathcal{X}', T=0)\big)\Big|$$

максимальна. Затем процедура рекурсивно повторяется для $\mathcal{X} \setminus \mathcal{X}'$.

На основе узорных структур

Узорные структуры

G - множество объектов, $(D, \sqcap)$ нижняя полурешетка всех возможных описаний объектов, и $\delta : G \to D$ отображение. Тройка $(G, (D, \sqcup), \delta)$называется узорной структурой, если $\forall X \subseteq \delta(G)$ $\bigsqcap X \in (D, \sqcap)$, где $\delta(G) = \{\delta(g) \mid g \in G\}$.

Элементы D называются **узорами** и частично упорядочены: $c \sqsubseteq d \Leftrightarrow c \sqcap d = c$, где $c, d \in D$.

Соответствие Галуа между G и $(D, \sqcap)$:

$$A^{\diamond} = \bigsqcap_{g \in A} \delta(g), \quad d^{\diamond} = \{g \in G \mid d \sqsubseteq \delta(g)\}, \quad \text{где } A \subseteq G, d \in D.$$

(A, d) называется узорным понятием, когда $A^{\diamond} = d$ и $d^{\diamond} = A$. $(\cdot)^{\diamond\diamond}$ является оператором замыкания на узорах.

Узорные структуры

Объект = пациент, узор = описание подгруппы.

a - количественный признак; b, c, d, e - количественные признаки;

$g_1, ..., g_6$ - объекты.

	a	b	c	d	e
g_1	$[1,2]$	1	1	1	0
g_2	$[3,3]$	1	1	1	1
g_3	$[2,3]$	1	1	1	1
g_4	$[2,2]$	0	1	0	0
g_5	$[3,4]$	1	0	1	1
g_6	$[4,5]$	1	1	0	1

$$a \in [1,4], \{b, c\}$$

$$\downarrow \diamond$$

$$\{g_1, g_2, g_3\}$$

$$\downarrow \diamond$$

$$a \in [1,3], \{b, c, d\}$$

Узорные структуры для выделения подгрупп

Преимущества

- Перебор только замкнутых описаний подгрупп позволяет существенно сократить полный перебор описаний в некотором классе описаний.
- Возможность отказа от жадного поиска, т.е. меньше шанс потерять какую-то важную подгруппу.
- Хорошая интерпретируемость описаний подгрупп.

Недостатки

- Экспоненциальная временная сложность построения узорной структуры.
- Большое число перебираемых подгрупп усложняет контроль ошибки неверного выделения хороших подгрупп.
- Проблемы с интепретацией пересекающихся подгрупп.
- Возможность возникновения парадокса Симпсона

Литература

Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment 2001; 5:1—56.

Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet 2005; 365:176—186.

Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. British Medical Journal 2010; 340:850—854.

Mayer C, Lipkovich I, Dmitrienko A. Survey results on industry practices and challenges in subgroup analysis in clinical trials. Statistics in Biopharmaceutical Research 2015; 7:272—282.

Lipkovich I, Dmitrienko A, D'Agostino Sr RB. Tutorial in Biostatistics: Data-Driven Subgroup Identification and Analysis in Clinical Trials. Statistics in Medicine 2016.

Xu Y, Yu M, Zhao YQ, Li Q, Wang S, Shao J. Regularized Outcome