Оценка эффекта множественного тестирования при поиске закономерностей в данных высокой размерности

Сенько О.В[1], Морозов А.М. [2], Кузнецова А.В. [3], Клименко Л.Л. [4]

1)ФИЦ ИУ РАН

2) ВМК МГУ им.М.В. Ломоносова

3) ИБХФ им. Н.М.Эмануэля РАН

4) ИХФ им. Н.Н.Семёнова РАН

Введение

- Одной из возможных технологий поиска таких закономерностей является метод оптимальных достоверных разбиений (ОДР), который использует для статистической верификации перестановочный тест. В условиях высокой размерности данных оценка достоверности двумерных закономерностей существенно осложняется проблемой множественного тестирования. Использование стандартного метода коррекции Бонферрони требует фиксации чрезвычайно жёстких и сильно завышенных порогов при отборе достоверных закономерностей при размерности данных выше 100.

Введение

Серия Монте-Карло экспериментов была проведена для оценки истинной достоверности закономерностей, выявленных при решении биомедицинской задачи изучения связи уровня фактора роста сосудов (VEGF) с широким набором биологических показателей. исследование основано на прямом подсчёте встречаемости двумерных закономерностей с различными нескорректированными уровнями значимости в общем наборе двумерных закономерностей, полученных с помощью метода ОДР.

Биомедицинская задача

Целью исследования было исследование взаимосвязи уровня содержания в сыворотке крови эндотелиальныого фактора роста кровеносных сосудов белка VEGF (vascular endothelial growth factor) с различнымми биологическими и биохимическими показателями.VEGF влияет на развитие новых кровеносных сосудов и выживание незрелых кровеносных сосудов.

База данных

- Изучалась связь VEGF с анамнезом, со стандартных биохимическими показателями, концентрацией гормонов щитовидной железы и половых гормонов, показателями коагуллограммы, концентрацией нейроспецифических белков, характеризующих повреждение мозговой ткани при ишемическом инсульте. В качестве X-переменных рассматривалась уровни макро- и микроэлементов в сыворотке крови, а также значения показателей энергетического метаболизма мозга Всего изучалась взаимосвязь целевой переменной со 142 показателями.

База данных

- В исследование была включена группа из 55 пациентов с возрастом от 40 до 88 лет, имеющих в анамнезе ишемический инсульт (ИИ) и группа из 33 пациентов с возрастом от 33 до 84 лет, имеющих в анамнезе случаи транзиторой ишемической атаки (ТИА).

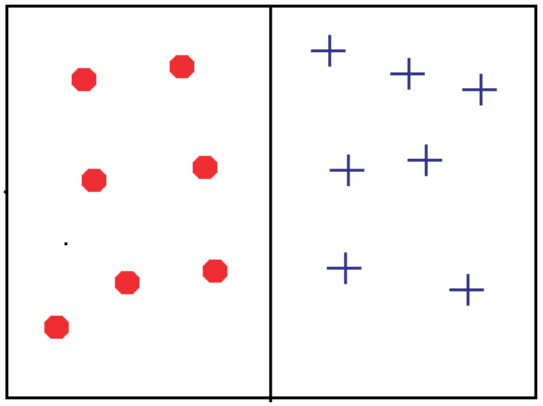
Метод оптимальных достоверных разбиений (ОДР)

Предположим, что нам требуется восстановить зависимость переменной Y от объясняющих переменных X_1, X_2 .

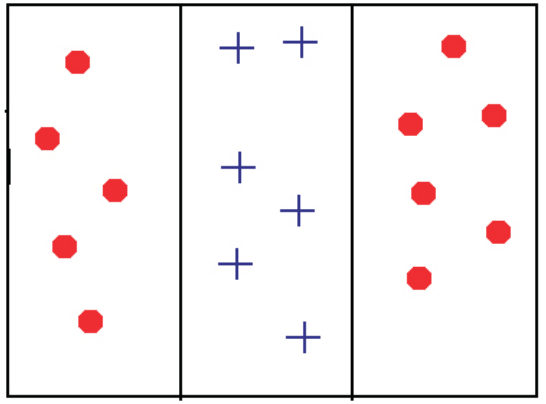
Для этого строятся разбиения области совместных допустимых значений объясняющих переменных, позволяющие наилучшим образом разделить объекты с различными уровнями значений переменной Y .

Одним из способом построения оптимальных разбиений является поиск таковых внутри семейств, имеющих фиксированную геометрическую форму.

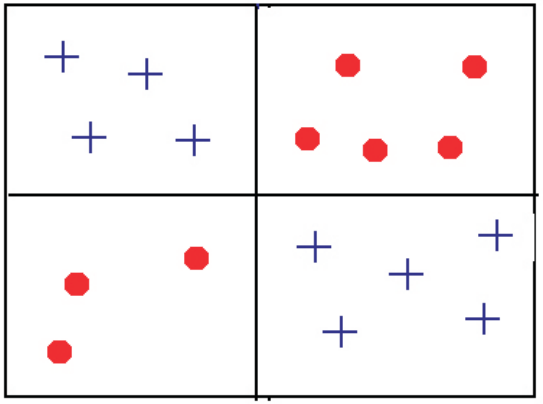
Модели разбиений с фиксированной геометрической формой



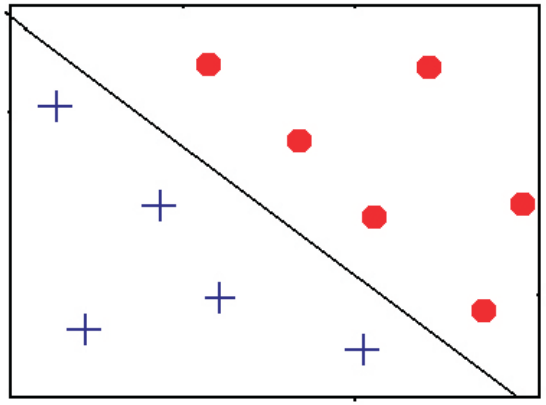
Family I



Family II



Family III



Family IV

Оптимизация разбиений

- Оптимизация разбиений производится по обучающей выборке

$$\tilde{S}_t = \{(y_1, \mathbf{x}_1 \ldots, (y_m, \mathbf{x}_m)\}$$

- и сводится к максимизации функционала

$$Q = \sum_{i=1}^{r} (\overline{y}_i - \overline{Y})^2 m_i$$

- m_i - число объектов элементе разбиения q_i

$\overline{y}_i$ - среднее значение Y в q_i

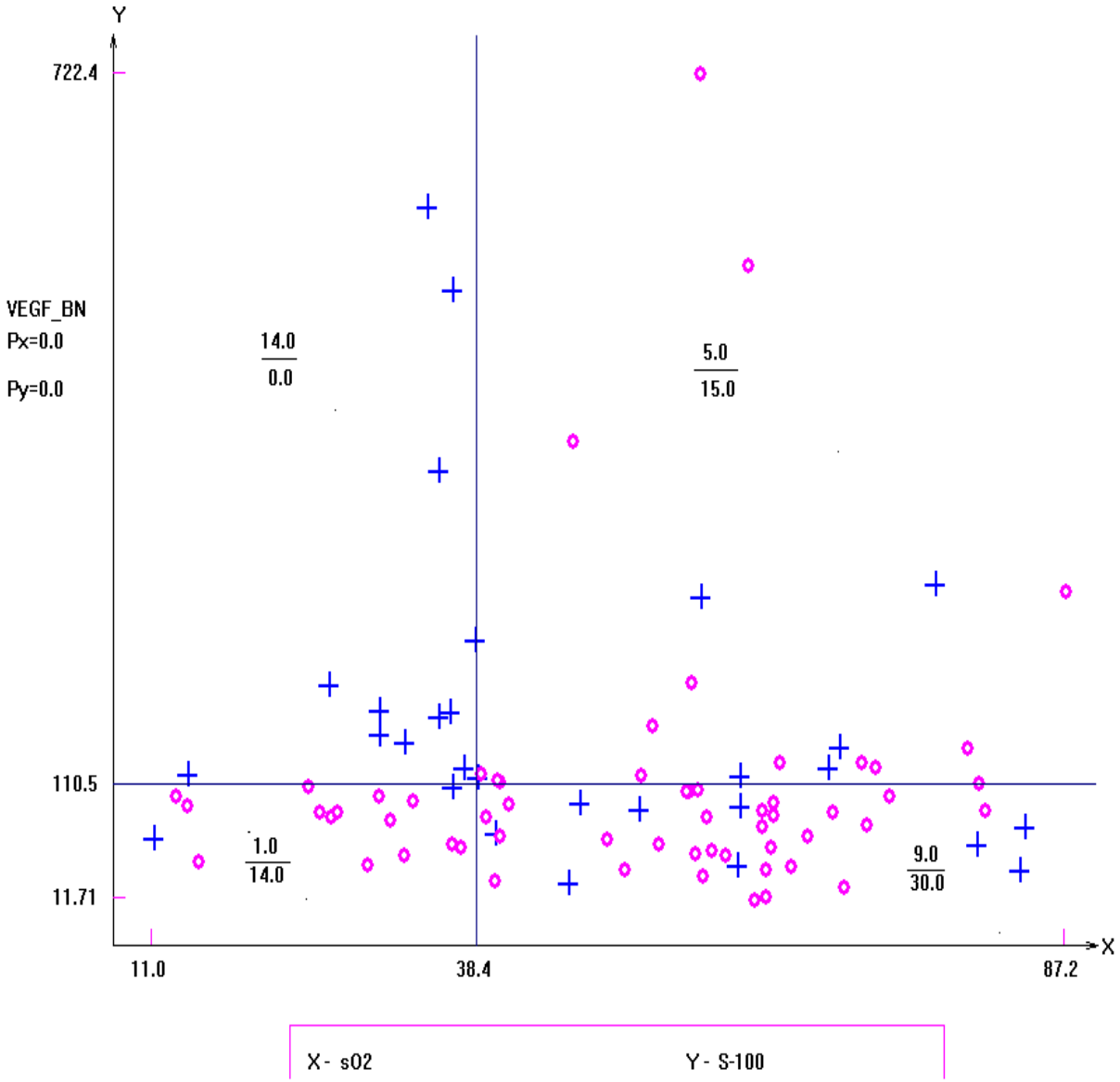
Y - среднее значение Y

Верификация закономерностей в методе ОДР

Верификация выявленных закономерностей оценивается через p-значений, которые вычисляется с помощью перестановочного теста (ПТ). Для моделей из семейств III оценивается значимость каждой из граничных точек (по осям X_1 и $X2$).

Значимость границы по оси X_1 (p1) оценивается с помощью случайных выборок, получаемых с помощью случайных перестановок Y внутри двух областей, задаваемых границей по оси X_2. Значимость границы по оси X_2 (p2) оценивается с помощью случайных выборок, получаемых с помощью случайных перестановок Y внутри двух областей, задаваемых границей по оси X_1.

Пример двумерной закономерности в методе ОДР



На рисунке представлена двумерная закономерность, связывающая бинарный показатель VEGF-bin с концентрации группы нейроспецифических белков S-100 и показателя насыщения (сатурации) крови кислородом (sO2)

Пример двумерной закономерности в методе ОДР

Случаи с уровнем VEGF выше 750 обозначеныы +,

случаи с уровнем VEGF ниже 750 обозначеныы o,

В каждом квадранте находится дробь, в числителе которой находится число случаев, обозначенных значком +, в знаменателе находится число случаев, обозначенных значком o.

Значимости по S-100 и sO2 описываются условиями

$P_1<0.0005$ $p2<0.0005$

Проблема множественного тестирования

Использовании двумерной модели метода ОДР для изучения зависимости Y от переменных из $\{X_1,\ldots,X_n\}$

сводится к проверке $\frac{1}{2}n(n-1)$ **пар нулевых гипотез о независимости Y от парных сочетаний переменных из**

$\{X_1,\ldots,X_n\}$

В силу самой природы статистической верификации вероятность случайного превышения значения статистики критерия для хотя бы одной нулевых гипотез из некоторого множества может быть существенно больше вероятности такого превышения при проверке одной индивидуальной гипотезы.

Проблема множественного тестирования

Использование для оценки истинной достоверности закономерностей известных методов коррекции Бонферрони, Бонферрони-Холма, Хохберга требует фиксации чрезвычайно жёстких и практически редко достижимых порогов при отборе достоверных закономерностей при размерности данных выше 100.

Эксперименты по оценке эффекта множественного тестирования

- Изучение эффекта множественного тестирования основывалось на сравнении индивидуальной достоверности закономерностей, найденных в случайных выборках, с индивидуальной достоверностью закономерностей, найденных в исходной выборке. При этом случайные выборки генерировались из исходной выборки путём случайных перестановок позиций значений целевой переменной относительно фиксированных позиций векторов X-переменных.

Эксперименты по оценке эффекта множественного тестирования

Таблица 1. Доли пар переменных, для которых выполняется условие $\max\{p_1, p_2\} \leqslant \alpha$.

α	ν	α	ν
$p < 0.0005$	0	0.007	$1.03 * 10^{-3}$
0.0005	$4.14 * 10^{-5}$	0.008	$1.2 * 10^{-3}$
0.001	$9.06 * 10^{-5}$	0.009	$1.5 * 10^{-3}$
0.0015	$1.38 * 10^{-4}$	0.01	$1.6 * 10^{-3}$
0.002	$2.2 * 10^{-4}$	0.012	$2 * 10^{-3}$
0.0025	$3 * 10^{-4}$	0.014	$2.4 * 10^{-3}$
0.003	$3.9 * 10^{-4}$	0.017	$3.07 * 10^{-3}$
0.0035	$4.57 * 10^{-4}$	0.02	$3.7 * 10^{-3}$
0.004	$5.16 * 10^{-4}$	0.025	$5 * 10^{-3}$
0.0045	$6.59 * 10^{-4}$	0.03	$6.2 * 10^{-3}$
0.0055	$8.33 * 10^{-4}$	0.05	$1.12 * 10^{-2}$

Эксперименты по оценке эффекта множественного тестирования

- Из таблицы 2 видно, что условие $\max(p1, p2) < 0.0005$
- не было достигнуто ни для одной пары X-переменных для всех 50 случайных выборок. Отсюда мы можем оценить на уровн е ниже 0.02 вероятность случайного возникновения конфигурации данных, которая соответствует существованию двумерной закономерности, удовлетворяющей условию

$$\max(p1, p2) < 0.0005$$

хотя бы одной пары X-переменных среди 10011 возможных пар.

СвязьVEGF с S-100 в сочетании с другими показателями

Таблица 2 Двумерные закономерности, в которых одним из факторов является S-100

Показ.	Границы	p—знач.	Распред.	
Hg	127.5	0.012	0/1	12/5
S-100	146.348	0.002	9/4	8/49
ОЖСС	39.5	0.002	5/10	17/20
S-100	86.738	0.002	6/0	1/29
pCO2	44.0	0.013	2/2	16/11
S-100	114.445	0.002	5/0	6/46
pO2	40.0	0.01	17/10	1/2
S-100	116.268	p<0.0005	6/47	5/0
sO2	38.4	p<0.0005	14/0	5/15
S-100	110.54	p<0.0005	1/14	9/30
FO2Hb	37.075	0.025	13/1	6/14
S-100	110.54	0.001	1/14	9/30
FHHb	54.6	0.007	3/11	15/1
S-100	116.268	p<0.0005	7/28	4/19
Ca	2.255	p<0.0005	2/16	11/2
S-100	114.44	0.007	28/68	18/5

Скорректированная достоверность для закономерностей из таблицы 2

- Следующей по уровню значимости в таблице 2 является двумерная закономерность, связывающая VEGF с показателями ОЖСС и S-100, для которой

$$\max(p1, p2) \leq 0.002$$

Вероятность случайного появления закономерности с такой индивидуальной достоверностью в совокупности из 10011 пар признаков оценивается с помощью таблицы 1 на уровне

$$1-(1-2.2*10^{-4})^{1011} \approx 0.89$$

Скорректированная достоверность для закономерностей из таблицы 2

Однако при ограничении исследования эффектами, проявляющимися в сочетании с S-100, вероятность случайного появления закономерности с индивидуальной достоверностью, удовлетворяющей условию $\max(p1, p2) \le 0.002$, среди 141 возможных парных сочетаний оценивается с помощью таблицы 1 на уровне

$$1-(1-2.2*10^{-4})^{141} \approx 0.0305 < 0.05$$

- Спасибо за внимание !!!