Оценка эффекта множественного тестирования в методе оптимальных достоверных разбиений*

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

senkoov@mail.ru, alxmopo3ov@gmail.com, azfor@narod.ru, klimenkoll@mail.ru

[1]ФИЦ «Информатика и управление» РАН, г. Москва, ул. Вавилова, 44/2

[2]МГУ им. М. В. Ломоносова, г. Москва, Ленинские горы, 1

[3]Институт биохимической физики им. Н. М. Эмануэля, г. Москва, ул. Косыгина, 4

[4]Институт химической физики им. Н. Н. Семёнова, г. Москва, ул. Косыгина, 4

Разработка методов поиска статистически достоверных эмпирических закономерностей является одной из приоритетных задач интеллектуального анализа данных. 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 жестких и практически редко достижимых порогов при отборе достоверных закономерностей при размерности данных выше 100. Серия Монте-Карло экспериментов была проведена для оценки истинной достоверности закономерностей, выявленных при решении биомедицинской задачи изучения связи уровня фактора роста сосудов (VEGF — vascular endothelial growth factor) с широким набором биологических показателей. Набор закономерностей, найденных в исходной выборке, сравнивался с наборами закономерностей, найденных в 50 случайных выборках, полученных из исходной путем случайных перестановок значений целевой переменной. Эксперименты показали, что доля двумерных закономерностей, для которых исходная статистическая значимость, рассчитанная с помощью нескорректированного теста не хуже фиксированного уровня α, оказывается в 10–30 раз ниже величины αN_p, где N_p — число просмотренных пар объясняющих переменных. В статье также обсуждаются подходы, направленные на смягчение условий достоверности для закономерностей.

Ключевые слова: *закономерности; перестановочный тест; множественное тестирование*

DOI: 10.21469/22233792.2.1.03

1 Введение

Создание новых биофизических и биохимических методов исследования живых организмов привело к значительному росту числа показателей, заносимых в биомедицинские базы данных. Изучение взаимосвязи данных показателей потенциально может привести к обнаружению новых эмпирических закономерностей, важных для понимания функционирования биологических систем, а также при решении разнообразных задач диагностики и прогнозирования. Однако эффективность поиска действительно достоверных общих для всей генеральной совокупности закономерностей снижается из-за известной проблемы множественного тестирования, состоящей в случайном возникновении в задачах высокой размерности таких конфигураций данных, которые ошибочно верифицируются стандартными нескорректированными статистическими тестами как достоверные закономерности.

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-07-00819.

При этом вероятность хотя бы одного ошибочного объявления конфигурации данных закономерностью может значительно превышать уровень значимости, рассчитанный с помощью стандартного теста. Для того чтобы обеспечить исключение из последующего анализа ложных закономерностей, необходимо использование дополнительных более жестких критериев отбора. Наиболее известными методами модификации статистических критериев являются известная поправка Бонферрони, фактически предложенная в работе [1]. В последующие годы был разработан ряд дополнительных уточняющих критериев, включая критерии Бонферрони–Холма [2], Шидака [3], Хохберга [4]. Настоящая работа посвящена оценке величины эффекта множественного тестирования при поиске двумерных закономерностей с помощью метода ОДР [5–7].

Метод ОДР направлен на поиск одномерных или двумерных закономерностей, описывающих зависимость целевой величины Y от переменных, обозначаемых обычно буквой X, которые далее называются X-переменными. Достоверные двумерные закономерности в методе ОДР ищутся через построение оптимальных разбиений совместных областей допустимых значений для всевозможных пар X-переменных. Верификация закономерностей, характеризуемых оптимальными разбиениями, производится с помощью специального варианта перестановочного теста. Отметим, что технология верификации, основанная на перестановочных тестах, не требует априорных предположений о типе распределений, легко реализуется при произвольном виде статистик и, несмотря на высокую трудоемкость, получает все большее распространение [8–11]. Перестановочным тестом проверяются нулевые гипотезы о независимости Y от X-переменных. При этом общее число таких гипотез равно числу всевозможных пар X-переменных, которое оказывается чрезмерно большим для многих биомедицинских задач. Вследствие этого использование приведенных выше способов коррекции, основанных на оценивании сверху вероятности отклонения хотя бы одной из нулевых гипотез, приводит к неоправданно жестким критериям отбора, серьезно затрудняющим применение двумерного ОДР анализа уже при нескольких десятках X-переменных. Настоящее исследование основано на прямом подсчете встречаемости двумерных закономерностей с различными нескорректированными уровнями значимости в общем наборе двумерных закономерностей, полученных с помощью метода ОДР.

2 Метод оптимальных достоверных разбиений

Метод ОДР представляет собой метод анализа данных, позволяющий описать зависимость целевой переменной Y от некоторой переменной X, или от пары переменных X_1 и X_2 по выборке $\tilde{S}_t$ вида $\{(y_1, x_1), \ldots, (y_m, x_m)\}$ или $\{(y_1, x_{11}, x_{12}), \ldots, (y_m, x_{m1}, x_{m2})\}$, где y_j — значение целевой переменной Y на объекте s_j, а x_j, x_{j1} и x_{j2} — значения на объекте s_j переменных X, X_1 и X_2 соответственно, $j = 1, \ldots, m$. В основе метода лежит попытка построить такое разбиение интервала допустимых значений переменной X или совместной области допустимых значений переменных X_1 и X_2, чтобы объекты выборки $\tilde{S}_t$, принадлежащие разным элементам разбиения, по возможности сильнее отличались по уровням значений переменной Y.

Поиск оптимальных разбиений. Разбиения ищутся внутри нескольких семейств различного уровня сложности, включая

1) семейство I, состоящее из одномерных разбиений с одной граничной точкой;
2) семейство II, состоящее из одномерных разбиений с двумя граничными точками;
3) семейство III двумерных разбиений с двумя границами, параллельными координатным осям.

На рис. 1–3 приведены примеры разбиений из каждого из трех упомянутых семейств.

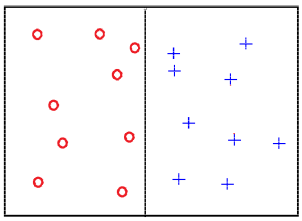

Рис. 1 Семейство I

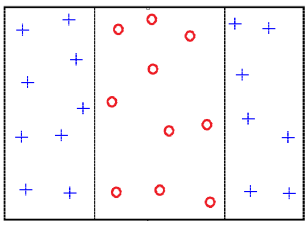

Рис. 2 Семейство II

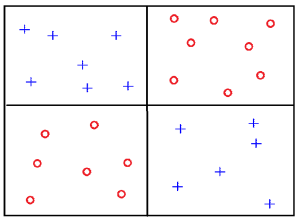

Рис. 3 Семейство III

Произвольное разбиение R из семейства I, которое далее будем обозначать $\widetilde{R}_I$, состоит из двух элементов (квадрантов) — q_1 и q_2. Произвольное разбиение R из семейства III, которое далее будем обозначать $\widetilde{R}_{III}$, состоит из четырех элементов (квадрантов) — q_1, q_2, q_3 и q_4. Пусть $\bar{Y}_0 = (1/m)\sum_{j=1}^{m} y_j$ — среднее значение целевой переменной по всей выборке $\widetilde{S}_t$; m_i — число объектов $\widetilde{S}_t$, для которых значения переменных X_1 и X_2 принадлежат квадранту q_i; $\bar{Y}_i$ — среднее значение целевой переменной по объектам $\widetilde{S}_t$, для которых значения переменных X_1 и X_2 принадлежат квадранту q_i; $D(Y)$ — дисперсия целевой переменной Y по всей обучающей выборке $\widetilde{S}_t$.

Оптимальным внутри семейства $\widetilde{R}_1$ считается разбиение, для которого достигает максимума значение функционала

$$Q^1(\widetilde{S}_t, R) = \frac{1}{D(Y)} \sum_{i=1}^{2} (\bar{Y}_0 - \bar{Y}_i)^2 m_i \,.$$

Оптимальным внутри семейства $\widetilde{R}_3$ считается разбиение, для которого достигает максимума значение функционала

$$Q^3(\widetilde{S}_t, R) = \frac{1}{D(Y)} \sum_{i=1}^{4} (\bar{Y}_0 - \bar{Y}_i)^2 m_i \,.$$

Разбиение, на котором достигается максимум функционалов Q^1 или Q^3, будет обозначаться R_o.

Верификация закономерностей. Верификация закономерностей из семейства $\widetilde{R}_1$ основана на попытке опровержения простой нулевой гипотезы о том, что целевая переменная Y не зависит от переменной X. Для этих целей используется вариант перестановочного теста, состоящий в многократном повторении поиска оптимального разбиения на множестве выборок $\{\widetilde{S}_t^f | f \in \widetilde{f}\}$, полученных из исходной выборки путем случайных перестановок значений целевой переменной Y относительно фиксированных значений переменной X. Через $\widetilde{f}$ обозначено получаемое с помощью генератора случайных чисел множество случайных перестановок чисел из набора $\{1, \ldots, m\}$. В качестве p-значения используется вероятность превышения величины $Q^1(\widetilde{S}_t, R_o)$ при условии соблюдения нулевой гипотезы. Данная вероятность оценивается как доля выборок из $\{\widetilde{S}_t^f | f \in \widetilde{f}\}$, для которых выполняется неравенство

$$Q^1(\widetilde{S}_t^r, R_o^f) \geqslant Q^1(\widetilde{S}_t, R_o).$$

Через R_o^f обозначено оптимальное разбиение, построенное по случайной выборке $\widetilde{S}_t^f$. Описанный вариант перестановочного теста, исследованный в работе [5], далее будем называть первым вариантом.

Первый вариант перестановочного теста не может быть использован для верификации закономерностей, задаваемых разбиениями из семейства III, поскольку его применение приводит к появлению на выходе большого числа так называемых частично ложных закономерностей. Под частично ложной понимается такая двумерная закономерность, для которой достоверность наличия связи между Y и двумя X-переменными на самом деле обеспечивается только одной переменной из пары X_1 и X_2. Включение же второй переменной является по сути случайным.

Второй вариант перестановочного теста основан на попытке опровержения нулевой гипотезы о достаточности одних только одномерных моделей для описания существующей связи. Подобный подход может трактоваться как вариант известного методологического принципа бритвы Оккама. На практике изучается возможность опровержения нулевых гипотез о достаточности одномерных разбиений, ближайших к верифицируемому двумерному разбиению. В качестве ближайших выступают одномерные разбиения, имеющие границы, совпадающие с соответствующими границами верифицируемого двумерного разбиения [7]. Нулевая гипотеза о достаточности одномерного разбиения R считается эквивалентной гипотезе о независимости Y от X-переменных внутри двух квадрантов R. Второй вариант перестановочного теста основан на проверке таких нулевых гипотез.

Опишем данный вариант подробно. Предположим, что оптимальное двумерное разбиение R_o^2 задается границей b_1 для переменной X_1 и границей b_2 для переменной X_2. Пусть R_1 и R_2 — одномерные разбиения, задаваемые границами b_1 и b_2 соответственно. Обозначим через $\widetilde{f}_R$ множество случайных перестановок чисел из набора $\{1,\ldots,m\}$ с запрещенным обменом номерами объектов из $\widetilde{S}_t$ с X-описаниями, принадлежащим разным квадрантам простого одномерного разбиения R. Сгенерируем с помощью датчика случайных чисел множества выборок $\widetilde{\mathbf{S}}_1 = \{\widetilde{S}_t^f | f \in \widetilde{f}_{R_1}\}$ и $\widetilde{\mathbf{S}}_2 = \{\widetilde{S}_t^f | f \in \widetilde{f}_{R_2}\}$. Второй вариант перестановочного теста вычисляет для R^2 два параметра:

1) p_1 — оценку вероятности достижения (или превышения) величины $Q^3(\widetilde{S}_t, R_o)$ при условии соблюдения нулевой гипотезы о независимости Y от X-переменных внутри двух квадрантов R_1;
2) p_2 — оценку вероятности достижения (или превышения) величины $Q^3(\widetilde{S}_t, R_o)$ при условии соблюдения нулевой гипотезы о независимости Y от X-переменных внутри двух квадрантов R_2.

Параметр p_i оценивается как доля выборок из $\widetilde{\mathbf{S}}_i$, для которых выполняется неравенство:

$$Q^3(\widetilde{S}_t^f, R_o^f) \geqslant Q^3(\widetilde{S}_t, R_o). \tag{1}$$

Рассчитанные по сгенерированным множествам выборок $\widetilde{\mathbf{S}}_1$ и $\widetilde{\mathbf{S}}_2$ с использованием неравенства (1) параметры p_1 и p_2 выступают в качестве p-значений. Будем считать, что параметр p_1 описывает достоверность опровержения нулевой гипотезы о достаточности R_2 для описания взаимосвязи Y с X-переменными и, следовательно, характеризует достоверность необходимости использования в двумерной закономерности переменной X_1.

Параметр p_2 описывает достоверность опровержения нулевой гипотезы о достаточности R_1 для описания взаимосвязи Y с X-переменными и характеризует достоверность необходимости использования в двумерной закономерности переменной X_2.

Двумерная закономерность считается значимой на уровне α, если одновременно выполняются неравенства $p_1 < \alpha$ и $p_2 < \alpha$. Использование изложенного подхода в методе ОДР описано в работе [6]. В работе [12] рассматривалось аналогичное применение пере-

становочных тестов для оценки необходимости аппроксимации данных кусочно-линейной регрессионной моделью вместо более простой линейной.

3 Проблема множественного тестирования при поиске закономерностей

Метод ОДР эффективно оценивает достоверность закономерности, связывающей целевую переменную Y с сочетанием переменных X_1 и X_2. Важную информацию для изучения зависимости переменной Y от совокупности переменных $\widetilde{X} = \{X_1, :, X_n\}$ может дать анализ двумерных закономерностей, связывающих Y со всевозможными парными сочетаниями переменных из $\widetilde{X}$. Поиск таких закономерностей сводится, согласно содержанию предыдущего раздела, к проверке набора нулевых гипотез о независимости Y от переменной из $\widetilde{X}$ внутри квадрантов простых разбиений. В силу самой природы статистической верификации вероятность случайного достижения (или превышения) значения статистики критерия для хотя бы одной нулевых гипотез из $H_1, \ldots, H_r$ значительно больше вероятности такого события при проверке одной индивидуальной гипотезы. В случае ОДР вероятность случайного достижения функционала качества $Q^3(\widetilde{S}_t, R_0)$ хотя бы для одного из парных сочетаний переменных из $\widetilde{X}$ может значительно превышать p-значения, рассчитанные при верификации отдельной закономерности без учета эффекта множественного тестирования. Вследствие этого настоящий уровень значимости найденной закономерности оказывается хуже уровня значимости, рассчитанного с помощью простого применения перестановочного теста. Проблему необходимости использования существенно более жестких критериев отбора при тестировании большого числа исходных нулевых гипотез принято называть проблемой множественного тестирования (множественных сравнений).

Наиболее распространенными способами оценивания верхних границ для вероятности случайного отклонения хотя бы одной нулевой гипотезы является метод коррекции Бонферрони–Холма. В данном методе коррекция уровня значимости производится путем простого умножения исходного уровня значимости α, рассчитанного с помощью используемого нескорректированного статистического критерия C, на множитель $r - r_v + 1$, где r — общее число проверяемых нулевых гипотез; r_v — общее число проверяемых нулевых гипотез, которые были отвергнуты на уровне значимости α. Иными словами, при наличии r_v нулевых гипотез, отвергнутых C на уровне α, эти гипотезы следует считать достоверно отвергнутыми на уровне $\alpha(r - r_v + 1)$. При использовании двумерных моделей типа III из метода ОДР общее число проверяемых нулевых гипотез очевидно равно удвоенному значению различных пар X-переменных или $r = n(n-1)$ (см. разд. 1). В современных биомедицинских базах данных общее число разнообразных клинических, лабораторных или инструментальных показателей, которые могут рассматриваться в качестве X-переменных, нередко достигает 150–200 или даже более высоких значений. Таким образом, общее число тестируемых нулевых гипотез достигает $2 \cdot 10^4$–$4 \cdot 10^4$, а величин множителя $r - r_v + 1$ может существенно превышать 10^4. Для того чтобы закономерности можно было достоверно считать значимыми на уровне $p < 0{,}05$ или $p < 0{,}01$ согласно методу Бонферрони–Холма, необходимо, чтобы p-значения, рассчитанные согласно способу из разд. 2 не превышали 10^{-6}. Для корректной оценки столь низких p-значений требуется свыше 10^6 перестановок, что потребовало бы чрезвычайно высоких объемов вычислений. Кроме того, столь высокая значимость закономерностей достигается редко при наиболее распространенных объемах баз клинических данных, включающих порядка 10^2–10^3 случаев. Однако теория Бонферрони–Холма основана на завышенной оценке вероятности

ошибочного отклонения нулевых гипотез в условиях множественного тестирования, что, в свою очередь, приводит к существенному занижению уровня достоверности выявляемых закономерностей. Существенно более точную картину может дать использование методов, основанных на сравнении наборов закономерностей, найденных в реальной выборке с наборами закономерностей, найденных в случайных выборках, имеющих сходные с исходной выборкой структуру и объем. Одним из способов генерации случайных выборок может быть случайная перестановка позиций целевой переменной относительно фиксированных позиций X-переменных. Это означает, что для целей коррекции, связанной с проблемой множественного тестирования, может быть использована фактически та же самая схема, которая лежит в основе перестановочных тестов, используемых для тестирования отдельных закономерностей. Следует отметить, что перестановочный тест достаточно активно используется для оценки величины эффекта множественного тестирования. В этой связи могут быть упомянуты работы [13, 14].

4 Задача анализа связи VEGF с другими биологическими показателями

Целью исследования было исследование взаимосвязи уровня содержания в сыворотке крови эндотелиальныого фактора роста кровеносных сосудов белка VEGF с различнымми биологическими и биохимическими показателями. VEGF влияет на развитие новых кровеносных сосудов (ангиогенез) и развитие незрелых кровеносных сосудов (сосудистая поддержка), запуская сигнальный каскад, который в конечном итоге стимулирует рост эндотелиальных клеток сосуда, их функционирование и пролиферацию [15]. Для достижения более высокой устойчивости и наглядности анализа на предварительном этапе непрерывный показатель содержания VEGF в сыворотке крови переводился в бинарную форму, т. е. в качестве целевой переменной использовался бинарный показатель VEGF-bin, равный 1 при содержании VEGF менее 750 нг/л и равный 2 при содержании VEGF более 750 нг/л.

Метод ОДР использовался для изучения связи VEGF со стандартными биохимическими показателями, концентрацией гормонов щитовидной железы и половых гормонов, показателями коагуллограммы, концентрацией нейроспецифических белков, характеризующих повреждение мозговой ткани при ишемическом инсульте (ИИ). В качестве X-переменных рассматривались также уровни макро- и микроэлементов в сыворотке крови, а также значения показателей энергетического метаболизма мозга — уровня постоянного потенциала (УПП). В общем, изучалась взаимосвязь целевой переменной со 142 показателями.

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

Использование изложенного выше метода ОДР со вторым вариантом престановочного теста при использовании 2000 случайных престановок выявило следующее распределение закономерностей, описываемых разбиениями из $\widetilde{R}_3$, по уровню значимости:

– 158 двумерных закономерностей, для которых

$$\max(p_1, p_2) < 0{,}05;$$

– 24 двумерные закономерности, для которых

$$\max(p_1, p_2) < 0{,}01;$$

– 12 двумерных закономерностей, для которых

$$\max(p_1, p_2) < 0{,}005;$$

– 1 двумерная закономерность, для которой

$$\max(p_1, p_2) < 0{,}0005.$$

Таким образом, одна из найденных двумерных закономерностей имеет согласно второму варианту перестановочного теста значимость, определяемую неравествами $p_1 < 0{,}0005$ и $p_2 < 0{,}0005$. Данная закономерность связывает бинарный показатель VEGF-bin с концентрацией нейроспецифических белков S-100 и показателем насыщения (сатурации) крови кислородом (sO2). Белки S-100 — группа кальцийсвязывающих белков с низким молекулярным весом, участвующих в регуляции разнообразных внутриклеточных и межклеточных процессов. Известно, что уровень S-100 коррелирует с повреждением мозговой ткани при ИИ. Также в ранней фазе церебрального инфаркта S-100 является ответом мозговой ткани на ишемию [16]. Индекс сатурации sO2 представляет собой долю гемоглобина, связанного с кислородом, и является важным показателем, характеризующим снабжение тканей кислородом [17]. Закономерность графически представлена на рис. 4.

Случаи с уровнем VEGF выше 750 обозначены +, случаи с уровнем VEGF ниже 750 обозначены **O**. В каждом квадранте находится дробь, в числителе которой находится число случаев, обозначенных значком +, в знаменателе находится число случаев, обозначенных значком **O**. Квадранты пронумерованы римскими цифрами с возрастанием номера по

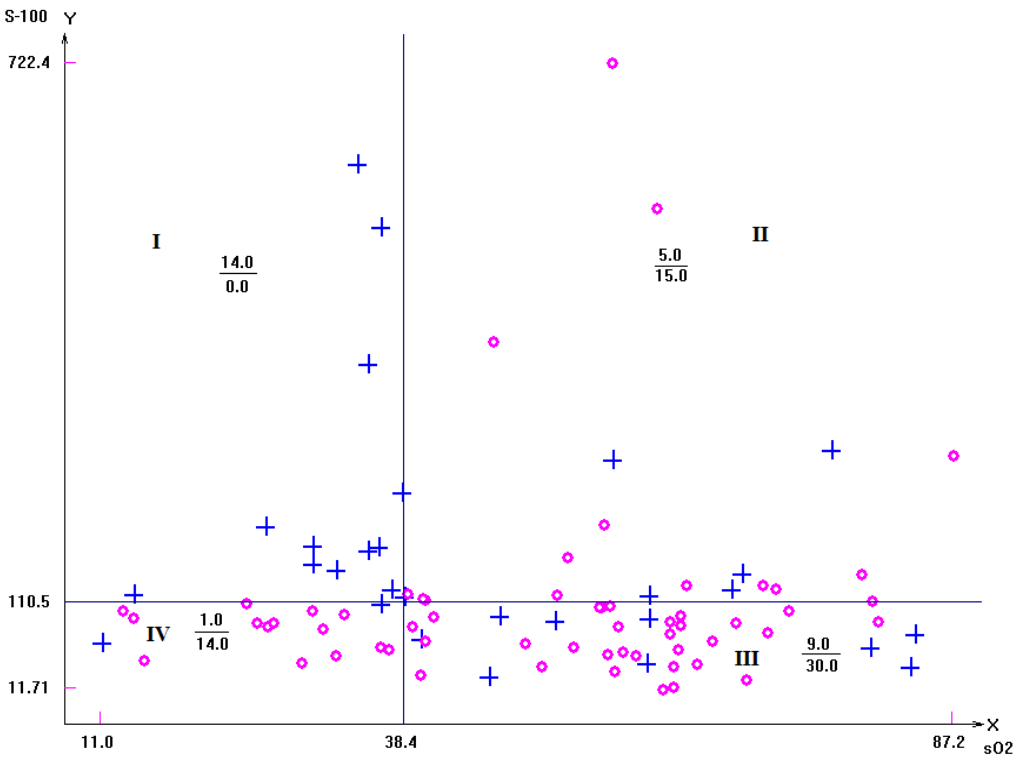

Рис. 4 Двумерная закономерность, связывающая коэффициент сатурации sO2 и S-100 с VEGF

ходу часовой стрелки. Нумерация начинается от верхнего левого квадранта. В квадранте I находятся только наблюдения с уровнем VEGF выше 750. В остальных квадрантах преобладают наблюдения из группы с низкими значениями VEGF. Наиболее сильное преобладание наблюдается в квадранте IV. Таким образом, низкий уровень сатурации sO2 в сочетании с высоким значением S-100 в преобладающем большинстве случаев соответствует высокому значению VEGF, что, возможно, связано с необходимостью компенсации недостаточного уровня снабжения головного мозга кислородом. Отметим, что связь содержания S-100 с VEGF выявляется также с использованием простейшей одномерной модели метода ОДР при $p = 0{,}002$.

Отметим, что метод ОДР позволил также выявить целый ряд двумерных закономерностей, описывающих связь между VEGF и S-100 в сочетании с целым рядом других показателей. В их число вошли общий уровень гемоглобина, фракции оксигемоглобина FO2Hb и дезоксигемоглобина FHHb в общем гемоглобине, общая железосвязывающая способность сыворотки (ОЖСС), парциальное давление кислорода в венозной крови (pO2), парциальное давление углекислого газа в венозной крови (pCO2), общее содержание Ca.

В ячейках табл. 1 для каждой закономерности даны значения вошедших в нее показателей, соответствующие границы и p-значения, рассчитанные с помощью второго варианта перестановочного теста. В двух правых колонках, озаглавленных «Распределение»,

Таблица 1 Двумерные закономерности, в которых одним из факторов является S-100

Показатели	Границы	p-значения	Распределение	
			VERF $>$ 750	VERF $<$ 750
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

представлено распределение случаев с VEGF > 750 и VEGF < 750. Дроби, приведенные в этих ячейках, имеют тот же смысл, что и дроби в квадрантах на рис. 4. Расположение ячеек в двух правых колонках таблицы совпадает с расположением соответствующих квадрантов.

Необходимо отметить, что почти все показатели, вошедшие в закономерности из табл. 1, непосредственно связаны со снабжением кислородом головного мозга.

5 Компьютерные эксперименты по оценке эффекта множественного тестирования

Изучение эффекта множественного тестирования основывается на сравнении достоверности закономерностей, найденных в случайных выборках, с достоверностью закономерностей, найденных в исходной выборке. При этом случайные выборки генерировались из исходной выборки путем случайных перестановок позиций значений целевой переменной относительно фиксированных позиций векторов X-переменных. Для оценивания достоверности двумерных закономерностей в случайных выборках использовался второй вариант перестановочного теста, т. е. для каждой закономерности вычислялись p-значения p_1 и p_2. Из-за высокой трудоемкости вычислений исследование ограничивалось 50 случайными выборками. Для набора уровней значимости α из отрезка $[0, 0{,}05]$ была рассчитана усредненная по всем 50 выборкам доля пар переменных, для которых были выявлены двумерные закономерности, удовлетворяющие условию $\max\{p_1, p_2\} \leqslant \alpha$. Указанные доли приведены в табл. 2. Значения уровней значимости приведены в столбцах таблицы, озаглавленных α. Соответствующая доля пар переменных приведена в соседнем столбце, озаглавленном ν. В верхней левой ячейке приведена доля пар переменных, удовлетворяющих условию $\max\{p_1, p_2\} < 0{,}0005$. Доли ν являются несмещенными и состоятельными оценками вероятности выполнения неравенства $\max\{p_1, p_2\} \leqslant \alpha$ (или неравенства $\max\{p_1, p_2\} < 0{,}0005$ для верхней левой ячейки) при выполнении условия независимости целевой переменной от вектора X-переменных. Из табл. 2 видно, что усредненная по всем 50 выборкам доля пар X-переменных, для которых выполнено условие

$$\max(p_1, p_2) < 0{,}0005\,, \tag{2}$$

составляет $1{,}18 \cdot 10^{-5}$. Используя данную долю в качестве оценки вероятности выполнения условия (2), получаем вероятность случайного появления хотя бы одной двумерной

Таблица 2 Доли пар переменных, для которых выполняется условие $\max\{p_1, p_2\} \leqslant \alpha$

α	ν	α	ν
$p < 0{,}0005$	$1{,}18 \cdot 10^{-5}$	0,007	$4{,}63 \cdot 10^{-4}$
0,0005	$3{,}35 \cdot 10^{-5}$	0,008	$5{,}57 \cdot 10^{-4}$
0,001	$5{,}71 \cdot 10^{-5}$	0,009	$6{,}28 \cdot 10^{-4}$
0,0015	$8{,}86 \cdot 10^{-5}$	0,01	$7{,}42 \cdot 10^{-4}$
0,002	$1{,}18 \cdot 10^{-4}$	0,012	$9{,}25 \cdot 10^{-4}$
0,0025	$1{,}52 \cdot 10^{-4}$	0,014	$1{,}18 \cdot 10^{-3}$
0,003	$1{,}77 \cdot 10^{-4}$	0,017	$1{,}53 \cdot 10^{-3}$
0,0035	$2{,}1 \cdot 10^{-4}$	0,02	$1{,}98 \cdot 10^{-3}$
0,004	$2{,}48 \cdot 10^{-4}$	0,025	$2{,}68 \cdot 10^{-3}$
0,0045	$2{,}8 \cdot 10^{-4}$	0,03	$3{,}48 \cdot 10^{-3}$
0,0055	$3{,}6 \cdot 10^{-4}$	0,05	$7{,}07 \cdot 10^{-3}$

«закономерности» среди 10011 пар:

$$1 - (1 - 1{,}18 \cdot 10^{-5})^{10011} \simeq 0{,}165\,.$$

Таким образом, «закономерность», удовлетворяющая условию (1), может возникнуть чисто случайно, по крайней мере для одной из пар переменных с вероятностью примерно 16,5%.

Следующей по уровню значимости в табл. 1 является двумерная закономерность, связывающая бинарный показатель VEGF-bin с показателями ОЖСС и S-100. Из табл. 1 видно, что все 6 случаев с ОЖСС $< 39{,}5$ и S-100 $< 86{,}738$ соответствуют высокому уровню VEGF. Наоборот, из 30 случаев с ОЖСС $> 39{,}5$ и S-100 $< 86{,}74$ высокому уровню VEGF соответствует только один случай.

Из табл. 1 также видно, что для данной закономерности $\max(p_1, p_2) = 0{,}002$. Согласно табл. 2 доля пар переменных, для которых выполнено условие

$$\max(p_1, p_2) \leqslant 0{,}002\,, \tag{3}$$

составляет $1{,}18 \cdot 10^{-4}$. Используя данную долю в качестве оценки вероятности выполнения условия (1), получаем вероятность случайного появления хотя бы одной двумерной «закономерности» среди 10011 пар:

$$1 - (1 - 1{,}18 \cdot 10^{-4})^{10011} \simeq 0{,}89\,.$$

Таким образом, «закономерность», удовлетворяющая условию (3), может возникнуть чисто случайно, по крайней мере для одной из пар переменных с вероятностью примерно 89%. Поэтому простое выполнение условий (2) и (3) не является сколь-либо убедительным свидетельством наличия соответствующих закономерностей, обладающих обобщающей способностью, если проводить исследование по полной совокупности наблюдаемых переменных. В целом из анализа табл. 2 можно сделать вывод, что при всеобъемлющем разведывательном анализе данных с размерностью выше 142 нельзя рассчитывать на достоверность даже тех двумерных закономерностей, для которых выполнено условие (2).

Однако нередко интересы исследователей ограничиваются существенно более узкой задачей, сводящейся к оценке характера связи с целевой величиной только какой-то определенной группы показателей или определенного набора сочетаний показателей. Например, двумерные закономерности, представленные в табл. 2, соответствуют оценке влияния на уровень VEGF содержания белков из группы S-100 в сочетании с другими биохимическими и биологическими показателями. Общее число парных сочетаний такого типа очевидно составляет 141. Вероятность случайного появления хотя бы одной двумерной закономерности, удовлетворяющей условию (2) среди 141 пар, будем оценивать точно так же, как ранее оценивалась аналогичная вероятность для 10011 пар, т. е.

$$1 - (1 - 1{,}18 \cdot 10^{-5})^{141} \simeq 0{,}0017 < 0{,}002\,.$$

Таким образом закономерность, связывающую VEGF с S-100 в сочетании с SO2, можно считать значимой на уровне $p < 0{,}02$ после проведения коррекции, учитывающей эффект множественного тестирования. Вероятность случайного появления хотя бы одной двумерной закономерности, удовлетворяющей условию (3) среди 141 пар, соответственно согласно таблице оценивается по формуле:

$$1 - (1 - 1{,}18 \cdot 10^{-4})^{141} \simeq 0{,}0165 < 0{,}02\,.$$

Закономерность, связывающую VEGF с S-100 в сочетании с ОЖСС, можно считать значимой на уровне $p < 0{,}02$ после проведения коррекции, учитывающей эффект множественного тестирования. К сожалению, учет эффекта множественного тестирования не позволяет сделать заключение о достоверности остальных закономерностей из табл. 2. Следует отметить, что простая коррекция по Бонферрони при размере множества пар переменных, в котором осуществляется поиск, равном 141, дает значимость закономерности, связывающей VEGF с S-100 в сочетании с sO2 всего лишь на уровне

$$p < 0{,}0005 \cdot 141 = 0{,}0705\,.$$

Значимость закономерности, связывающей VEGF с S-100 в сочетании с ОЖСС, оценивается на уровне

$$p = 0{,}002 \cdot 141 = 0{,}282\,.$$

Таким образом, обе закономерности оказываются незначимыми.

6 Заключение

Проведенные оценочные расчеты показывают, что двумерная закономерность, полученная с помощью метода ОДР и удовлетворяющая условию $\max(p_1, p_2) < 0{,}0005$, оказывается значимой на уровне $p < 0{,}002$, а двумерная закономерность, полученная с помощью метода ОДР и удовлетворяющая условию $\max(p_1, p_2) < 0{,}002$, оказывается значимой на уровне $p < 0{,}02$ с учетом эффекта множественного тестирования при переборе 141 пары переменных. При этом доля двумерных закономерностей, для которых исходная статистическая значимость, рассчитанная с помощью нескорректированного теста, не хуже фиксированного уровня α оказывается в 10–30 раз ниже величины αN_p, где N_p — число протестированных пар переменных. Для того чтобы анализ оставался достоверным для менее выраженных закономерностей, необходимо ограничивать число просматриваемых пар X-переменных, исходя из формулируемых на начальном этапе целей.

К биологическим результатам исследования следует отнести выявление взаимосвязи уровня белка VEGF с показателями, характеризующими снабжение тканей кислородом, которая, однако, проявляется только в сочетании с содержанием белков из группы S-100.

Литература

[1] *Dunn O. J.* Multiple comparisons among means // J. Am. Stat. Association, 1961. Vol. 56(293). P. 52–64.

[2] *Holm S.* A simple sequentially rejective multiple testprocedure // Scand. J. Stat., 1979. No. 6. P. 65–70.

[3] *Sidak Z. K.* Rectangular confidence regions for the means of multivariate normal distributions // J. Am. Stat. Association, 1967. No. 62(318). P. 626–633.

[4] *Hochberg Y.* A sharper Bonferroni procedure for multiple tests of significance // Biometrika, 1988. Vol. 75. P. 800–802.

[5] *Сенько О. В.* Перестановочный тест в методе оптимальных разбиений // Ж. вычисл. матем. матем. физ., 2003. Т. 43. № 9. С. 1422–1431.

[6] *Senko O., Kuznetsova A.* The optimal valid partitioning procedures // "InterStat" — Statistics in Internet, June 2006. No. 6. http://ip.statjournals.net.

[7] *Kuznetsova A., Kostomarova I., Senko O.* Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients // Pattern Recogn. Image Anal., 2013. Vol. 22. No. 4. P. 10–25.

[8] *Kim H.-J., Fay M. P., Feuer E. J., Midthune D. N.* Permutation tests for joint point regression with applications to cancer rates // Stat. Medicine, 2000. Vol. 19. No. 3. P. 335–351.

[9] *Ernst M.* Permutation methods: A basis for exact inference // Stat. Sci., 2004. Vol. 19. No. 4. P. 676–685.

[10] *Good P. I.* Permutation, parametric and bootstrap tests of hypotheses. — Springer ser. in statistics. — 3rd ed. — Springer, 2005. 334 p.

[11] *Ojala M., Garriga G.* Permutation tests for studying classifier performance // J. Machine Learning Res., 2010. No. 11. P. 1833–1863.

[12] *Senko O. V., Dzyba D. S., Pigarova E. A., Rozhinskaya L. Ya., Kuznetsova A. V.* A method for evaluating validity of piecewise-linear models // KDIR, 2014. P. 437–443.

[13] *Tusher V. G., Tibshirani R., Chu G.* Significance analysis of microarrays applied to the ionizing radiation response // Proc. Natl. Acad. Sci. USA, 2001. Vol. 98. P. 5116–5121.

[14] *Dudoit S., Popper Shaffer J., Boldrick J. C.* Multiple hypothesis testing in microarray experiments // Stat. Sci., 2003. Vol. 18. No. 1. P. 71–103.

[15] *Sun Y., Jin K., Xie L., Childs J., Mao X. O., Logvinova A., Greenberg D. A.* VEGF-induced neuroprotection, neurogenesis, and angiogenesis after focal cerebral ischemia // J. Clin. Invest., 2003. Vol. 111. No. 12. P. 1843–1851, 976.

[16] *Marenholz I., Heizmann C. W., Fritz G.* S100 proteins in mouse and man: From evolution to function and pathology (including an update of the nomenclature) // Biochem. Biophys. Res. Commun., 2004. Vol. 322. No. 4. P. 1111–1122. doi:10.1016/j.bbrc.2004.07.096. PMID 15336958

[17] *Haymond S.* Oxygen saturation // Clinical Laboratory News, February 2006. No. 10-12. www.aacc.org.

Поступила в редакцию 15.10.2015

Evaluating of multiple testing effect in method of optimal valid partitioning*

***O. V. Senko*[1], *A. M. Morozov*[2], *A. V. Kuznetsova*[3], *and L. L. Klimenko*[4]**

senkoov@mail.ru, alxmopo3ov@gmail.com, azfor@narod.ru, klimenkoll@mail.ru

[1]Federal Research Center "Computer Science and Control" of RAS, 44/2 Vavilova st., Moscow, Russia

[2]Lomonosov Moscow State University, 1 Leninskie Gory, Moscow, Russia

[3]Emanuel Institute of Biochemical Physics RAS, 4 Kosygina st., Moscow, Russia

[4]Semenov Institute of Chemical Physics RAS, 4 Kosygina st., Moscow, Russia

Development of methods for statistically valid regularities discovery is one of the most important data mining problems. One of the possible techniques of regularities search is method of optimal valid partitioning (OVP), using permutation test for statistical verification. In high-dimensional tasks, verification becomes more complicated task due to the problem of multiple testing. Standard Bonferroni correction is based on very strong validity thresholds that rarely are practically achievable when dimension is greater than 100. Set of Monte-Carlo experiments was conducted to evaluate true validity of found regularities in the following biomedical task: study of relationship between vessels growth factor (VEGF) levels and wide set of biological indicators. Set of regularities found in initial data set was compared with sets of regularities

*The research was supported by the Russian Foundation for Basic Research, project No. 14-07-00819.

that were found in 50 random data sets. At that random data sets were generated from initial data set by random permutations of the target variable positions with fixed positions of explanatory variables vectors. It was shown in experiments that fraction of two-dimensional regularities that are valid at uncorrected significance level α is 10-30 times less than αN_p where N_p is the number of enumerated pairs of explanatory variables. Some ways to soft validity thresholds are discussed.

Keywords: *regularities; permutation test; multiple comparing*

DOI: 10.21469/22233792.2.1.03

References

[1] Dunn, O. J. 1961. Multiple comparisons among means. *J. Am. Stat. Association* 56(293):52–64.

[2] Holm, S. 1979. A simple sequentially rejective multiple testprocedure. *Scand. J. Stat.* 6:65–70.

[3] Sidak, Z. K. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Association* 62(318):626–633.

[4] Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802.

[5] Senko, O. V. 2003. Perestanovochnyi test v metode optimalnych razbienii. *Zh. Vychisl. Matem. Matem. Fiz.* 43(9):1422–1431.

[6] Senko, O., and A. Kuznetsova. June 2006. The optimal valid partitioning procedures. *"InterStat" — Statistics in Internet* 6. http://ip.statjournals.net.

[7] Kuznetsova, A., I. Kostomarova, and O. Senko. 2013. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. *Pattern Recogn. Image Anal.* 22(4):10–25.

[8] Kim, H.-,J., M. P. Fay, E. J. Feuer, and D. N. Midthune. 2000. Permutation tests for joint point regression with applications to cancer rates. *Stat. Medicine* 19(3):335–351.

[9] Ernst, M. 2004. Permutation methods: A basis for exact inference. *Stat. Sci.* 19(4):676–685.

[10] Good, P. I. 2005. *Permutation, parametric and bootstrap tests of hypotheses*. Springer ser. in statistics. 3rd ed. Springer. 334 p.

[11] Ojala, M., and G. Garriga. 2010. Permutation tests for studying classifier performance. *J. Machine Learning Res.* 11:1833–1863.

[12] Senko, O. V., D. S. Dzyba, E. A. Pigarova, L. Ya. Rozhinskaya, and A. V. Kuznetsova. 2014. A method for evaluating validity of piecewise-linear models. *KDIR* 437–443.

[13] Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116–5121.

[14] Dudoit, S., J. Popper Shaffer, and J. C. Boldrick. 2003. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* 18(1):71–103.

[15] Sun, Y., K. Jin, L. Xie, J. Childs, X. Mao, A. Logvinova, and D. A. Greenberg. 2003. VEGF-induced neuroprotection, neurogenesis, and angiogenesis after focal cerebral ischemia. *J. Clin. Invest.* 111(12):1843–1851, 976.

[16] Marenholz, I., C. .W. Heizmann, and G. Fritz. 2004. S100 proteins in mouse and man: From evolution to function and pathology (including an update of the nomenclature). *Biochem. Biophys. Res. Commun.* 322(4):1111–1122. doi:10.1016/j.bbrc.2004.07.096. PMID 15336958

[17] Haymond, S. February 2006. Oxygen saturation. *Clinical Laboratory News* 10-12. www.aacc.org.

Received October 15, 2015