Московский государственный университет имени М. В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Кодрян Максим Станиславович

Метод поиска статистически достоверных отклонений характеристик регрессионных зависимостей в различных областях признакового пространства

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф-м.н., в.н.с.

О. В. Сенько

Москва, 2018

Содержание

1 Введение

Задача поиска закономерностей является одной из фундаментальных в сфере интеллектуального анализа данных. Построение любого решающего правила в задачах регрессии и классификации в первую очередь опирается на способность выявлять различные зависимости, присутствующие в данных. Всякое решение задач кластеризации и анализа отклонений, без сомнения, основывается на возможности нахождения ассоциаций и связей в структуре данных. Большинство задач машинного обучения нуждаются в некотором инструменте, позволяющем в том или ином виде находить закономерности в данных по конечной выборке из генеральной совокупности объектов.

Однако и сама по себе задача поиска и, главное, верификации закономерностей имеет немалую значимость, например, в сфере биоинформатики и доказательной медицины. Часто не так важно определить конкретный вид связи, сколько оценить её значимость — степень уверенности в том, что найденная связь не фиктивна (случайна), а действительно присутствует в генеральной совокупности. Особое внимание данной проблеме уделяется в областях биологии, медицины, химии и так далее, поскольку данные в этих сферах часто имеют весьма разнообразную структуру с сравнительно высоким числом признаков и малым количеством объектов (анализы пациентов, уровни экспрессии генов на ДНК-микрочипах, результаты проведения химических реакций и т. п.), потому построение адекватного решающего правила в таких случаях бывает весьма проблематичным и при этом зачастую не столь насущным. Куда важнее выявить в данных те факторы, которые, взаимодействуя, образуют интересующие исследователя связи (например, поиск именно тех показателей в анализе крови, которые действительно связаны с наличием изучаемого заболевания), причём наличие данных связей необходимо доказать (с той или иной степенью «уверенности»).

Решению задачи поиска и верификации закономерностей в данных посвящена весомая часть исследований в областях математической статистики, машинного обучения, прикладной алгебры и так далее. Было создано огромное количество разнообразных методов: от проверки статистических гипотез до построения байесовских сетей и анализа формальных понятий. Данная работа же концентрируется на решении упомянутой задачи именно с точки зрения математической статистики. Таким образом, под «нахождением закономерности» в данных подразумевается проверка некоторой статистической гипотезы об отсутствии зависимости заданного вида (т. е. опровержение подобной гипотезы эквивалентно подтверждению наличия закономерности). Такой подход является хорошо исследованным и весьма неплохо зарекомендовавшим себя в связи, например, с тем, что позволяет придать термину «значимость закономерности» весьма интерпретируемую форму — вероятность случайного получения выборки из генеральной совокупности, в структуре которой прослеживается связь подобной «силы». Примером использования статистического подхода для поиска и верификации закономерностей на практике является статья [1], где описывается статистическая процедура анализа экспрессии генов на ДНК-микрочипах.

Данная дипломная работа представляет собой развитие предыдущей курсовой работы [2], посвящённой описанию, реализации и модификации метода Оптимальных Достоверных Разбиений (ОДР) [3–5], а также применению его для выявления и оценки статистической значимости закономерностей в медицинских данных в условиях множественной проверки гипотез. Была произведена модификация метода, позволяющая существенно расширить класс потенциально выявляемых закономерностей. В итоге был получен готовый программный инструмент для нахождения и статистической верификации связей различного вида, который был тщательно протестирован и применён на реальных медицинских данных для подтверждения наличия определённой зависимости между некоторыми показателями в биохимических анализах человека. Полученый результат оказался весьма интересным.

Название данной ВКР отражает суть основной реализованной модификации метода ОДР — способность находить и подтверждать условно-линейные связи между различными факторами — однако не ограничивает результаты проделанной работы, которые будут подробно описаны далее.

2 Постановка задачи

Основной целью данной работы являлась разработка модификаций метода ОДР, расширение области его применимости, устранение некоторых недостатков, а также проведение экспериментов на действительных медицинских данных для подтверждения наличия определённой (условно-линейной) связи между белками VEGF и S-100 относительно группы показателей оксиметрии (уровней насыщенности крови кислородом).

Стоит упомянуть, что биологи давно заметили наличие эмпирической связи между вышеупомянутыми факторами [6–10]. Особенно интересен характер данной связи в период восстановления пациента после ишемического инсульта [8]. Были проведены исследования, подтверждающие, что уровень пептида VEGF связан с гипоксемией (т. е. с определёнными значениями показателей оксиметрии) [10], однако до сих пор весомых доказательств наличия нетривиальной зависимости VEGF–S-100–оксиметрия получено не было: ни стандартные статистические тесты, ни корреляционный анализ удовлетворительных результатов не дали. К примеру, рассчитанный для построенной модели линейной регрессии, предсказывающей по уровням показателей оксиметрии и белка S-100 значение целевого признака VEGF, коэффициент детерминации принял значение $R^2 \approx 0.14$, что явно свидетельствует о недостаточности простой линейной модели для описания наблюдаемой зависимости.

Одна из попыток исследовать данную закономерность была описана в статье [11], однако в ней использовалась старая версия метода ОДР, в связи с чем, например, пришлось бинаризовать целевой признак VEGF (с непрерывным ни одной статистически значимой закономерности найти не удалось), что несомненно привело к искажению характера реальной зависимости в данных; также не было возможности оценить вклад группового эффекта сразу нескольких признаков в закономерность; в силу вычислительной неэффективности старой

реализации удалось получить лишь приблизительные результаты с весьма сомнительной теоретической гарантией корректности. Тем не менее, даже учитывая все перечисленные упрощения и недочёты, была установлена статистически верифицированная связь между белками VEGF, S-100 и одним из основных показателей оксиметрии sO2, что косвенно подтвердило эмпирическое предположение и послужило стимулом для дальнейших исследований.

Дальнейшее изложение сконцентрировано на описании предложенных и реализованных модификаций метода ОДР, позволивших решить данную практическую задачу (доказательство наличия и конкретизация вида связи меду VEGF, S-100 и показателями оксиметрии) и получить весьма многообещающие результаты.

2.1 Данные для исследования

Выборка данных, по которой производилось исследование использовалась та же, что и в статье [11], за исключением того факта, что целевой признак (VEGF) не подвергался бинаризации, а был оставлен в неизменном виде как веществennозначный показатель. Итак, данные представляли собой выборку из 88 пациентов: 55 пациентов с возрастом от 40 до 88 лет, имеющих в анамнезе ишемический инсульт (ИИ), и 33 пациента с возрастом от 33 до 84 лет, имеющих в анамнезе случаи транзиторой ишемической атаки (ТИА). Каждый объект (пациент) описывался при помощи веществennозначной целевой переменной VEGF и 142 биохимических показателей: концентрации гормонов щитовидной железы и половых гормонов, показатели коагуллограммы, концентрации нейроспецифических белков, характеризующих повреждение мозговой ткани при ишемическом инсульте (ИИ), уровни макро- и микроэлементов в сыворотке крови, значения показателей энергетического метаболизма мозга и прочие.

3 Модифицированный метод Оптимальных Разбиений

Метод Оптимальных Разбиений в своём первоначальном варианте используется для нахождения одномерных и двумерных закономерностей в данных путём поиска оптимального разбиения признакового пространства на квадранты с точки зрения значения некоторого функционала. Величина функционала характеризует «силу» найденной таким образом закономерности в топологии данных. Метод подробно описан в таких работах, как [2–5, 11]. В данной работе в силу особенностей исследуемых закономерностей (упор делается на линейную зависимость между целевой переменной и определённым признаком, начиная с некоторого порога по некоторому другому признаку), а также способа их исследования (рассматриваются закономерности относительно *группы* показателей оксиметрии) в исходный метод были внесены некоторые поправки, которые будут рассматриваться далее.

3.1 Поиск условно-линейных двумерных закономерностей

Эмпирическим путём была установлена сильная линейная связь между ответом VEGF и показателем S-100 у пациентов с определёнными значениями показателей оксиметрии (в особенности таких, как sO2 и FHHb) и отсутвие таковой при прочих значениях. Для того чтобы описывать закономерности подобного рода с точки зрения некоторой числовой характеристики, требуется несколько модифицировать классический метод оптимальных разбиений и ввести специальный функционал.

Итак, рассмотрим следующую выборку: $S = \{(x_i^1, x_i^2, y_i)\}_{i=1}^N$, состоящую из N объектов с двумя описательными и одним целевым признаками. Будем искать в ней закономерности, которые описываются следующим правилом (т. н. *условно-линейная зависимость*): «имеется существенная линейная зависимость между признаком x^1 и целевой переменной y для объектов со значением признака x^2, меньшим/большим некоторого порога b_{x^2}, при этом для остальных объектов характер линейной зависимости резко меняется (т. е. практически не наблюдается либо вовсе отсутствует)». Для этого введём следующий функционал:

$$Q(S, b_{x^2}) = \frac{m_l m_r \, ||\rho_l| - |\rho_r||}{1 - \max(\rho_l^2, \rho_r^2)}, \tag{1}$$

где m_l — количество объектов со значением признака x^2, меньшим порога b_{x^2}, m_r — с, соответственно, большим, ρ_l — значение робастного коэффициента корреляции Пирсона между целевой переменной y и признаком x^1, полученным на объектах со значением признака x^2, меньшим порога b_{x^2}, ρ_r — с, соответственно, большим.

Из вида формулы (1) становится ясно, что функционал принимает высокие значения именно в интересующих ситуациях: когда в одной из подвыборок, полученных из исходной выборки разделением по признаку x^2, наблюдается сильная линейная связь между признаком x^1 и ответом y, а в другой — наоборот, линейной зависимости не прослеживается. Причём упор также делается на приблизительно одинаковое соотношение объектов в каждой из подвыборок для получения состоятельных результатов: легко понять, что без домножения на количество объектов m_l и m_r (штрафуется несбалансированность по числу объектов) возможно возникновение ситуации, когда функционал примет высокое значение в связи с тем, что в одной из подвыборок объектов окажется совсем мало, а значит, высока вероятность получить **случайную** сильную линейную зависимость, — такие ситуации являются вырожденными и не представляют практического интереса.

Далее для данной выборки находится оптимальное значение порога $b_{x^2}^{opt}$ по следующему правилу:

$$b_{x^2}^{opt} = \underset{b_{x^2} \in [b_{x^2}^{left}, b_{x^2}^{right}]}{\operatorname{argmax}} Q(S, b_{x^2}),$$

где границы $b_{x^2}^{left}$ и $b_{x^2}^{right}$ подбираются таким образом, чтобы выполнялось условие «слева и справа (по признаку x^2) от границы b_{x^2} находится хотя бы n_{min} объектов». Значение функционала $Q(S, b_{x^2}^{opt})$ полагается оптимальным для данной выборки и далее обозначается как $Q(S)$.

В связи с наличием выбросов в выборке было принято решение в качестве ρ-показателя выбрать именно робастный коэффициент корреляции Пирсона. Вкратце, суть его получения заключается в следующем: по выборке строится модель линейной регрессии по паре признак-ответ (x^1, y). Далее из выборки удаляются объекты, для которых $|\hat{y} - y| > 3\sigma$, где $\hat{y}$ — предсказанное регрессионной моделью значение y по признаку x^1, σ — выборочное стандартное отклонение значений $\hat{y} - y$. Таким образом, из выборки будут удалены объекты, явно не вписывающиеся в модель, то есть выбросы. По отфильтрованной таким образом выборке подсчитывается классический коэффициент корреляции Пирсона. Данный коэффициент и называется робастным коэффициентом корреляции Пирсона, вычисленным по исходной выборке.

Упомянутый ранее параметр n_{min}, отвечающий за минимальное количество объектов, должное содержаться в левой и правой подвыборках, для данной задачи был положен равным 25. Данное значение было подобрано экспериментально и на практике оказалось оптимальным. Связано столь, на первый взгляд, высокое обязательное количество объектов в каждой подвыборке с неустойчивостью робастного коэффициента корреляции Пирсона при малом количестве наблюдений, что продемонстрировано на рисунке 1:

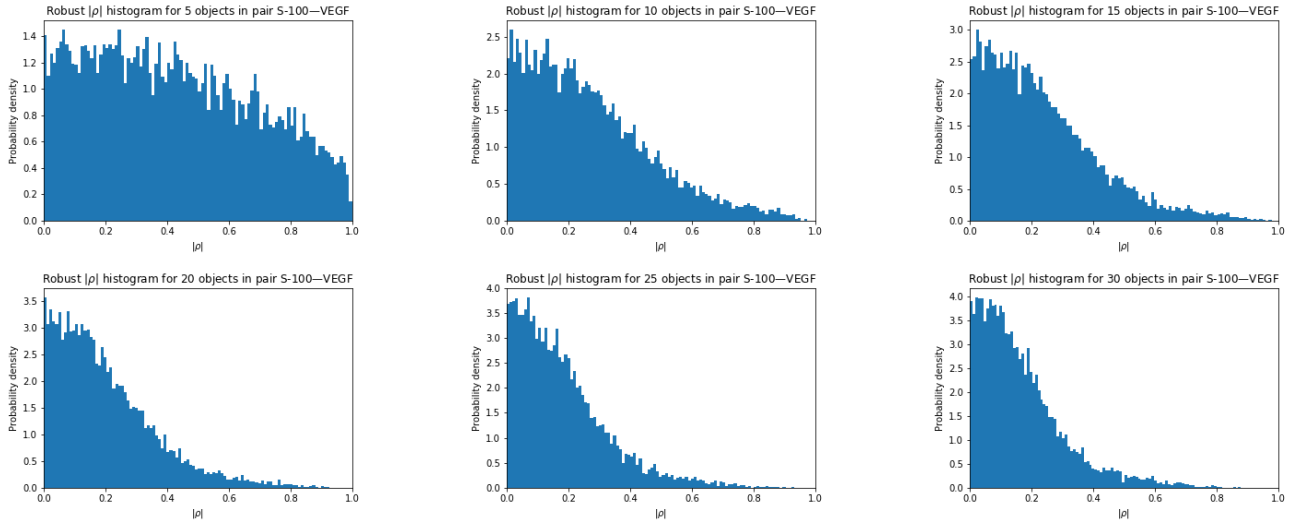


Рис. 1: Гистограммы плотности распределения модуля робастного коэффициента корреляции Пирсона (РККП) (обозначено как $|\rho|$) для выборок разного размера. Каждая гистограмма была получена путём рассчёта РККП для 10 000 выборок фиксированного размера, состоящих из пар случайно выбранных значений показателей S-100 и VEGF из исходной выборки данных. На оси абсцисс отложены значения модуля РККП, на оси ординат — значения плотности распределения.

4 Модифицированный метод Оптимальных Достоверных Разбиений

Обобщением метода оптимальных разбиений, позволяющим помимо приписывания найденным закономерностям некоторой числовой характеристики (например, оптимальное значение функционала) давать обоснованный ответ на вопрос о реальной статистической значимости каждой из найденных закономерностей, является так называемый метод Оптимальных Достоверных Разбиений (ОДР) [2–5, 11]. Для верификации найденных закономерностей метод использует аппарат проверки статистических гипотез. Таким образом, каждую выявляемую в данных закономерность можно отождествить с соответствующей нулевой гипотезой об отсутствии связи заданного вида (гипотеза о независимости). Процедура проверки подобных гипотез позволяет выделить действительно (статистически) значимые закономерности в данных.

В упомянутой ранее статье [11], являющейся по сути первой попыткой решить поставленную в начале данной работы задачу, проверяется набор гипотез вида «зависимости между целевым признаком (VEGF) и парой признаков f–g нет» для всевозможных различных признаков f и g.

В данной работе проверяемые гипотезы имеют следующий вид: «условно-линейной зависимости между целевым признаком (VEGF) и признаком f относительно группы показателей оксиметрии нет». Проверка каждой такой гипотезы сводится к проверке набора гипотез об отсутствии условно-линейной зависимости между ответом и признаком f относительно каждого показателя g из группы оксиметрии и некоторой аггрегации полученных результатов для принятия решения об отвержении исходной гипотезы.

В методе ОДР для проверки статистических гипотез используется техника перестановочного тестирования [3, 12, 13]. Опишем подробнее процедуру верификации двумерных закономерностей в ОДР для рассматриваемого случая.

4.1 Верификация двумерных закономерностей

Определим множество искусственных выборок $\mathcal{S}$. Каждый элемент этого множества — выборка, полученная из исходной выборки S путём случайного перемешивания значений её целевой переменной, т. е., если говорить формально,

$$\mathcal{S} = \left\{ \tilde{S} \mid |\tilde{S}| = |S| = N, \exists \sigma : S[i] = (x_i^1, x_i^2, y_i), \tilde{S}[i] = (x_i^1, x_i^2, y_{\sigma(i)}), i = 1, \ldots, N \right\},$$

где σ — некоторая перестановка множества $\{1, \ldots, N\}$, через $|S|$ обозначено количество объектов в выборке S, а через $S[i]$ — её i-ый элемент.

В каждой такой выборке в силу перемешивания значений ответа y теряется всякая присутствовшая связь между целевым и описательными признаками, а значит, всякая найденная

в подобной выборке закономерность между y и x будет иметь **случайную** природу, что даёт возможность судить о «степени случайности» закономерности в исходной выборке S.

Мощность множества $\mathcal{S}$ — $N!$, так как любая искусственно полученная выборка взаимно однозначно определяется соответствующей перестановкой σ. При разумных объёмах выборки это огромное число, поэтому мы не будем использовать все возможные искусственные выборки из множества $\mathcal{S}$, а выделим подмножество $\mathcal{S}_{func} \subset \mathcal{S}$. Мы используем данное подмножество искусственных выборок для оценки статистической значимости найденной закономерности в паре рассматриваемых признаков x^1, x^2 относительно целевой переменной y.

Для характеристики статистической значимости двумерной закономерности проведём процедуру рассчёта оптимального значения функционала (1) для каждой из выборок в $\mathcal{S}_{func}$ и введём следующие показатели:

$$p(S) = \frac{\left|\left\{\tilde{S} \in \mathcal{S}_{func} \mid Q(\tilde{S}) \geq Q(S)\right\}\right|}{|\mathcal{S}_{func}|},$$
$$h(S) = \frac{Q(S)}{\max_{\tilde{S} \in \mathcal{S}_{func}} Q(\tilde{S})}.$$

Как видно из определения, p-величина — доля тех случайных выборок, в которых двумерная закономерность оказалась «не хуже» той, что мы получили на исходной выборке, а h-величина — по сути то, во сколько раз исходная закономерность оказалась мощнее лучшей искусственной с точки зрения описанного функционала. Заметим, что обе величины получены при помощи перестановочного теста, так как для этого понадобились искусственно сгенерированные выборки. Величина p приблизительно равна вероятности случайно получить закономерность «не хуже» исходной при условии, что искомой зависимости нет. Величина h оказывается полезной для сравнения сильных закономерностей, у которых p-значение оказывается слишком малым.

4.2 Учёт группового эффекта

В рассматриваемой задаче основной целью было продемонстрировать и доказать наличие указанной выше (условно-линейной) зависимости между целевым показателем VEGF и признаком S-100 относительно группы показателей оксиметрии. Для того чтобы продемонстрировать данный эффект, требуется корректно расширить уже описанное понятие двумерной закономерности на случай, когда вторичных признаков, т. е. таких, по которым ищется граница разбиения выборки на две подвыборки (признак x^2), несколько.

В данной работе предлагается следующая модификация классического метода ОДР для учёта группового эффекта в многомерных зависимостях: использовать результаты верификации двумерных закономерностей на парах из рассматриваемого признака и отдельных признаков из группы для получения статистики о силе общей связи. Опишем данную идею более подробно.

Положим теперь, что выборка данных имеет следующий вид: $S = \{(x_i^1, x_i^2, \ldots, x_i^m, y_i)\}_{i=1}^N$, т. е. теперь каждый объект выборки — это m-мерный набор признаков и ответ. Обозначим через $G = \{g_1, \ldots, g_k\}$ некоторое подмножество $\{x^1, \ldots, x^m\}$ имеющихся признаков, которое будем называть *группой*. Нашей задачей является проверка наличия или отсутствия связи между данной группой, целевой переменной и каким-либо из остальных $m - k$ признаков.

Пусть f — какой-либо признак, не принадлежащий группе G. Разберём процесс получения сведений об имеющейся связи между группой G, ответом y и признаком f.

Первым этапом будет нахождение всех двумерных закономерностей в парах f–g_j, $j = 1, \ldots, k$. Для этого достаточно применить метод ОДР, описанный в начале данного раздела для каждой из выборок вида $S_j = \left\{(x_i^f, x_i^{g_j}, y_i)\right\}_{i=1}^N$. После этого мы получим набор p- и h-величин, характеризующих статистическую значимость закономерности в каждой паре f–g_j.

Чтобы оценить значимость совокупной закономерности между признаком и целевой переменной относительно группы G, введём следующие величины:

$$p_{f,G}(S) = median\left(p_{f,g_1}(S), \ldots, p_{f,g_k}(S)\right),$$
$$h_{f,G}(S) = median\left(h_{f,g_1}(S), \ldots, h_{f,g_k}(S)\right).$$

То есть групповая закономерность y–f–G теперь характеризуется двумя показателями $p_{f,G}$ и $h_{f,G}$, являющимися медианными усреднениями p- и h-значений, полученных для двумерных закономерностей вида y–f–g, где $g \in G$. Можно придать величинам $p_{f,G}(S)$ и $h_{f,G}(S)$ смысл некоторой числовой характеристики совокупной закономерности подобно тому, как это делалось для двумерных закономерностей с использованием оптимального значения функционала (1). Стоит отметить, что в качестве аггрегирующей функции была выбрана именно медиана по причине её устойчивости к отдельным выбросам (в отличие, например, от среднего значения).

Выделим некоторое подмножество $\mathcal{S}_{stats} \subset \mathcal{S}$ множества всех искусственных выборок $\mathcal{S}$, полученных из выборки S путём перемешивания значений целевого признака y. Проведём для каждой из искусственных выборок множества $\mathcal{S}_{stats}$ ту же самую процедуру. Теперь можно определить следующие величины:

$$P_{p_{f,G}}(S) = \frac{\left|\left\{\tilde{S} \in \mathcal{S}_{stats} \mid p_{f,G}(\tilde{S}) \leq p_{f,G}(S)\right\}\right|}{|\mathcal{S}_{stats}|},$$
$$P_{h_{f,G}}(S) = \frac{\left|\left\{\tilde{S} \in \mathcal{S}_{stats} \mid h_{f,G}(\tilde{S}) \geq h_{f,G}(S)\right\}\right|}{|\mathcal{S}_{stats}|},$$
$$H_{h_{f,G}}(S) = \frac{h_{f,G}(S)}{\max_{\tilde{S} \in \mathcal{S}_{stats}} h_{f,G}(\tilde{S})}.$$

Видно, что первые две величины имеют примерно тот же смысл, что и p-величина в случае двумерной закономерности — это вероятность случайно получить закономерность «не хуже» с точки зрения конкретной характеристики: $p_{f,G}$ или $h_{f,G}$. Обратим внимание, что

в зависимости от того, по какой из характеристик оценивается статистическая значимость совокупной закономерности, выбирается знак $\leq$ или $\geq$, так как чем ниже p-значение у закономерности, тем она сильней (для h — наоборот). Последняя величина по аналогии может быть интерпретирована как h-значение, введённое ранее.

4.3 Решение проблемы множественного тестирования

При проверке большого числа статистических гипотез исследователь зачастую сталкивается с *эффектом множественного тестирования* (ЭМТ) [14–19]. Суть данного эффекта заключается в резком повышении вероятности случайно отвергнуть нулевую гипотезу (совершить ошибку первого рода), если количество проверяемых гипотез велико.

В качестве примера рассмотрим простую теоретическую задачу. Пусть вероятность наступления некоторого события равна 0.01. Найдём вероятность того, что при проведении 100 независимых испытаний хотя бы раз наступит данное событие. Несложно понять, что ответом будет являться следующее число: $1 - (1 - 0.01)^{100} \approx 0.634$. Видно, насколько сильно возросла вероятность столкнуться с данным, на первый взгляд, маловероятным событием.

На практике ЭМТ вызывает много проблем, так как подсчитанные обычными методами, не учитывающими данный эффект, уровни значимости оказываются совершенно неточными. Для решения данной проблемы были разработаны различные методы коррекции найденных уровней значимости (например, поправка Бонферрони [16,19], где уровень значимости просто домножается на число проверяемых гипотез, Холма [15, 17, 19] — классическая нисходящая процедура — и прочие), однако практически все они основываются на завышенной оценке исправленных величин, поэтому оказываются малоприменимыми в задачах с действительно большим числом проверяемых гипотез. В связи с этим нарастает популярность менее консервативных методов оценки ЭМТ, основывающихся на упомянутом ранее перестановочном тестировании [18].

В рассмотренной задаче поиска закономерности относительно некоторой группы признаков также приходится иметь дело с проблемой множественного тестирования, так как для каждого из $m - k$ оставшихся признаков проверяется гипотеза о том, что «связи между данным признаком, ответом и группой нет». Рассмотрим вариант оценки ЭМТ в данной задаче при помощи перестановочного тестирования.

Снова выделим некоторое подмножество случайных выборок $\mathcal{S}_{mul} \subset \mathcal{S}$. Для каждой из них проведём все те же процедуры нахождения закономерностей между отдельными признаками и ответом относительно фиксированной группы. Для экономии вычислительных ресурсов мы будем использовать полученные ранее (при поиске закономерностей в исходной выборке) значения необходимых статистик для искусственных выборок из $\mathcal{S}_{func}$ и $\mathcal{S}_{stats}$. Данная процедура корректна в силу независимости используемых случайных выборок.

Итак, совершив необходимые процедуры, мы будем иметь набор величин $\left\{P_{p_{f,G}}(\tilde{S}), P_{h_{f,G}}(\tilde{S}), H_{h_{f,G}}(\tilde{S}) \mid \tilde{S} \in \mathcal{S}_{mul}, f \in \{x^1, \ldots, x^m\} \setminus G\right\}$. Ясно, что полученные ве-

личины имеют полностью случайный характер, так как каждая из выборок множества $\mathcal{S}_{mul}$ является искусственно сгенерированной и не содержит в себе никаких обусловленных закономерностей. Теперь зафиксируем некоторый уровень значимости α и получим оценку вероятности события «случайно (то есть при условии, что связи между *всеми* признаками и ответом нет) получить хотя бы одну закономерность между каким-либо из признаков и ответом относительно группы G не хуже уровня значимости α по показателю $P_{p_{f,G}}$»:

$$P_{mul}(P_{p_{f,G}} \mid \alpha) = \frac{\left|\left\{\tilde{S} \in \mathcal{S}_{mul} \mid \exists f \in \{x^1, \ldots, x^m\} \setminus G : P_{p_{f,G}}(\tilde{S}) \leq \alpha\right\}\right|}{|\mathcal{S}_{mul}|}. \tag{2}$$

По сути вероятность интересующего события была оценена как доля случайных выборок, у которых хотя бы для одного признака наблюдалась закономерность сильнее фиксированного уровня значимости α. Если быть точным, то это соответствует *слабому контролю над метрикой FWER* в теории проверки множественных гипотез [18, 19]. Такой подход является менее консервативным, чем, например, поправка Бонферрони, и при этом учитывает имеющиеся зависимости между проверяемыми гипотезами: действительно, признаки в данных имеют собственную внутреннюю взаимосвязанную структуру, что обязательно следует учитывать, проверяя гипотезы о наличии закономерностей между ответом и признаками.

В полной аналогии определим величину $P_{mul}(P_{h_{f,G}} \mid \alpha)$, являющуюся оценкой вероятности аналогичного события, но уже относительно показателя $P_{h_{f,G}}$. Определяя подобную величину для показателя $H_{h_{f,G}}$, учтём, что чем выше h-значение, тем сильнее связь, а значит, уровень значимости $\alpha \approx 0$ заменим уровнем значимости $\beta \approx 1$, а знак $\leq$ в (2) заменим на $\geq$.

В итоге получим три формулы для ЭМТ-коррекции уровня значимости: $P_{mul}(P_{p_{f,G}} \mid \alpha)$, $P_{mul}(P_{h_{f,G}} \mid \alpha)$ и $P_{mul}(H_{h_{f,G}} \mid \beta)$.

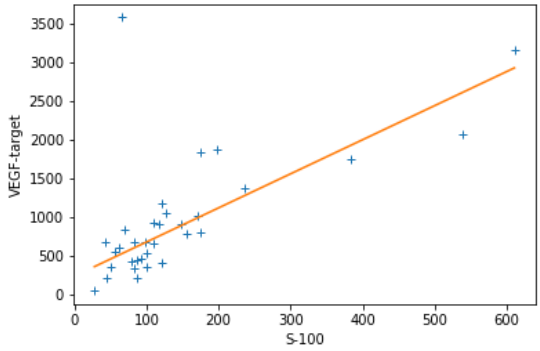
Теперь для оценки ЭМТ можно пользоваться следующим правилом: для рассматриваемой пары признак–группа в исходной выборке выбрать интересующий показатель ($P_{p_{f,G}}$, $P_{h_{f,G}}$ или $H_{h_{f,G}}$); вероятность случайного получения закономерности с таким показателем (без учёта ЭМТ) подставить вместо уровня значимости α в соответствующую формулу вида (2); полученное значение интерпретировать как скорректированную вероятность с учётом ЭМТ.

Обратим внимание, что ранее (например, в статье [11]) использовалась иная процедура для оценки эффекта множественного тестирования. Суть её заключалась в аппроксимации среднего числа ложно отвергаемых гипотез в случайных выборках на уровне значимости α, что в дальнейшем использовалось для ЭМТ-оценки. Такая процедура обладает рядом недостатков (например, не учитывает зависимости в проверяемых гипотезах) и является существенно менее теоретически обоснованной по сравнению с новой, описанной в данном разделе (подробнее в Приложении В). Возможность перехода к новому способу оценки ЭМТ обеспечивает независимость случайных выборок в множествах $\mathcal{S}_{mul}$, $\mathcal{S}_{stats}$ и $\mathcal{S}_{func}$: в предыдущей версии метода они оказывались зависимыми, потому для каждой случайной выборки из $\mathcal{S}_{mul}$ приходилось строить свои множества $\mathcal{S}_{stats}$ и $\mathcal{S}_{func}$, что, разумеется, накладывает существенные ограничения на объём $\mathcal{S}_{mul}$, а следовательно, и на точность получаемых оценок.

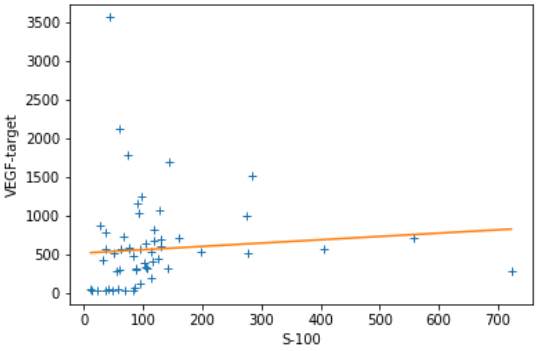
5 Результаты

Как было упомянуто ранее, в поставленной задаче нас более всего интересует доказательство наличия характерной линейной зависимости между ответом VEGF и показателем S-100 у пациентов с определёнными уровнями оксиметрии (группа показателей sO2, FHHb, FO2Hb) и отсутствия таковой у прочих пациентов.

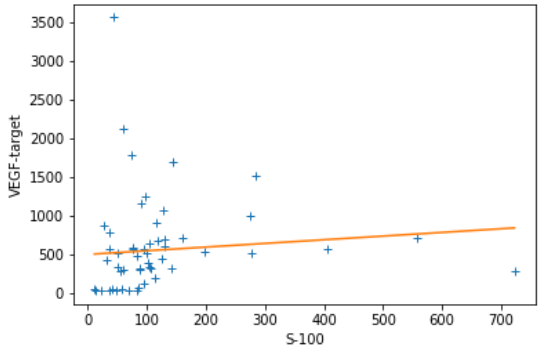
Итак, пусть далее $G = \{\mathrm{sO2}, \mathrm{FHHb}, \mathrm{FO2Hb}\}$. Для перестановочных тестов используем подмножества $\mathcal{S}_{func}, \mathcal{S}_{stats}, \mathcal{S}_{mul}$ случайных перестановок выборки S мощности 10 000 каждое. Применив описанный выше метод к предоставленной выборке и исследовав зависимости относительно показателя S-100, удалось получить следующие результаты:



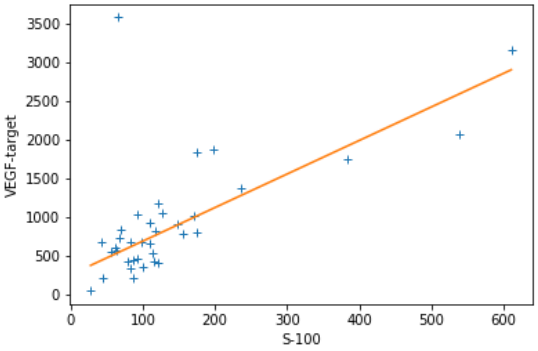
sO2 < 39.75: $\rho_l \approx 0.88$



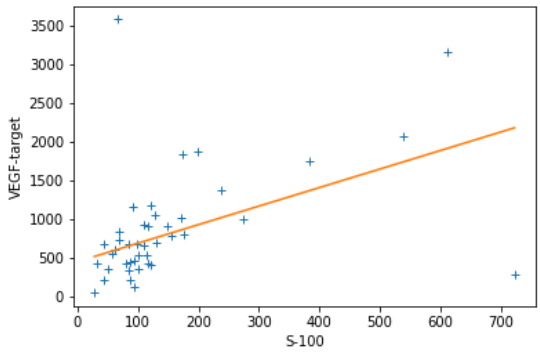
sO2 ≥ 39.75: $\rho_r \approx 0.12$



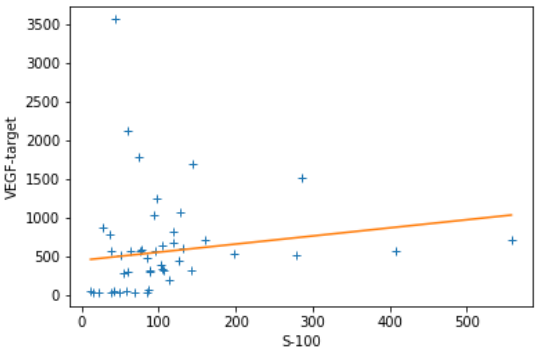
FHHb < 56.3: $\rho_l \approx 0.13$



FHHb ≥ 56.3: $\rho_r \approx 0.87$



FO2Hb < 44.45: $\rho_l \approx 0.59$



FO2Hb ≥ 44.45: $\rho_r \approx 0.21$

Рис. 2: Иллюстрация наличия явной линейной зависимости в паре VEGF–S-100, начиная с определённого порога для каждого из трёх исследованных показателей оксиметрии. Под каждым рисунком указан конкретный показатель оксиметрии, порог, по которому образуется подвыборка и значение робастного коэффициента корреляции Пирсона, расчитанного на данной подвыборке.

Таблица 1: Таблица результатов, полученных при исследовании групповой закономерности между целевым показателем VEGF, S-100 и группой показателей оксиметрии sO2, FHHb, FO2Hb.

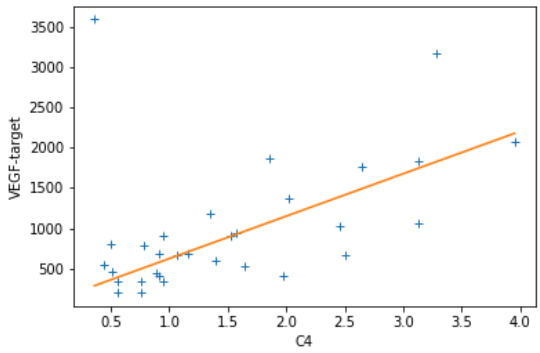
f	S-100
G	sO2, FHHb, FO2Hb
$p_{f,G}(S)$	0.0006
$h_{f,G}(S)$	0.63
$P_{p_{f,G}}(S)$	0.0001
$P_{h_{f,G}}(S)$	0.0
$H_{h_{f,G}}(S)$	1.016

Таблица 2: Таблица скорректированных уровней значимости для величин $P_{p_{f,G}}$, $P_{h_{f,G}}$ и $H_{h_{f,G}}$ для минимального из исследованных α, равного 10^{-4}, и различных β.

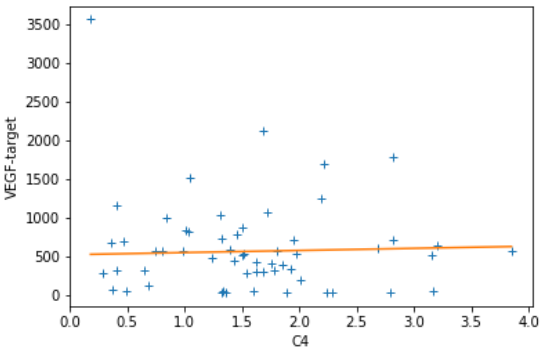
$P_{mul}(P_{p_{f,G}} \mid \alpha = 0.0001)$	0.02
$P_{mul}(P_{h_{f,G}} \mid \alpha < 0.0001)$	0.012
$P_{mul}(H_{h_{f,G}} \mid \beta = 1)$	0.012
$P_{mul}(H_{h_{f,G}} \mid \beta = 1.05)$	0.0082
$P_{mul}(H_{h_{f,G}} \mid \beta = 1.1)$	0.0066

Полученные результаты доказывают наличие условной линейной связи между VEGF и S-100 относительно показателей оксиметрии. Даже с учётом эффекта множественного тестирования, уровень значимости полученной закономерности не превышает 2% по наиболее консервативному из показателей — $P_{p_{f,G}}$. Действительно, заметим, что в таблице 2, представлены скорректированные уровни значимости (величины P_{mul}) для минимального из исследованных α, равного 10^{-4}, при котором скорректированное значение с учётом ЭМТ не превосходит 0.02. При этом полученная закономерность имеет (нескорректированную) значимость на уровне $\approx 1 \cdot 10^{-4}$, поэтому, исходя из вышесказанного, итоговый уровень значимости даже с учётом поправки не может превысить 0.02. Интересно, что в данном случае даже наиболее консервативная поправка Бонферрони оказалась бы менее строгой ($142 \cdot 10^{-4} = 0.0142 < 0.02$), что является еще одним аргументом в пользу подобного оценивания ЭМТ: используемая поправка в целом не столь сильно «штрафует» уровни значимости, но при этом более «подозрительно» относится к чрезмерно низким их значениям. По прочим показателям ($P_{h_{f,G}}$ и $H_{h_{f,G}}$) ситуация ещё более благоприятная: с учётом ЭМТ получаем скорректированный уровень значимости ≈ 0.012.

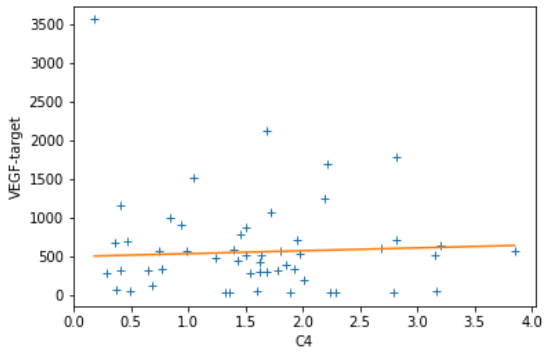
Более того, удалось обнаружить существенно сильную зависимость между целевым признаком VEGF и фактором C4 относительно данной группы показателей оксиметрии, что продемонстрировано на рисунке 3 и в таблице 3.



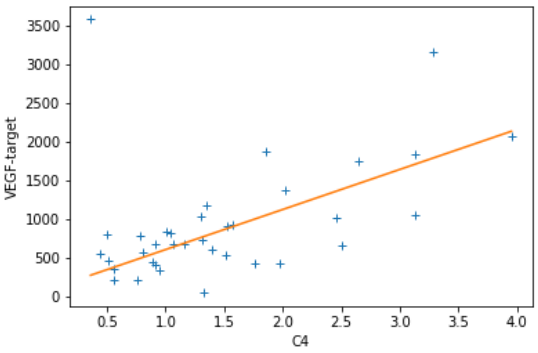
sO2 < 39.25: $\rho_l \approx 0.76$



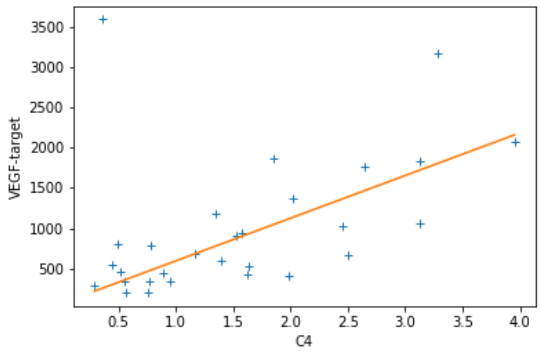
sO2 ≥ 39.25: $\rho_r \approx 0.05$



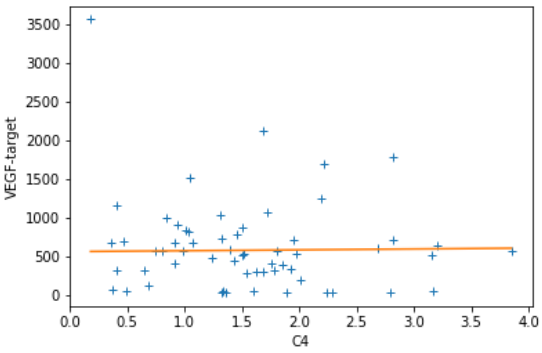
FHHb < 56.3: $\rho_l \approx 0.07$



FHHb ≥ 56.3: $\rho_r \approx 0.73$



FO2Hb < 37.075: $\rho_l \approx 0.76$



FO2Hb ≥ 37.075: $\rho_r \approx 0.02$

Рис. 3: Иллюстрация наличия явной линейной зависимости в паре VEGF–C4, начиная с определённого порога для каждого из трёх исследованных показателей оксиметрии. Под каждым рисунком указан конкретный показатель оксиметрии, порог, по которому образуется подвыборка и значение робастного коэффициента корреляции Пирсона, расчитанного на данной подвыборке.

Таблица 3: Таблица результатов, полученных при исследовании групповой закономерности между целевым показателем VEGF, C4 и группой показателей оксиметрии sO2, FHHb, FO2Hb.

f	C4
G	sO2, FHHb, FO2Hb
$p_{f,G}(S)$	0.0001
$h_{f,G}(S)$	0.9745
$P_{p_{f,G}}(S)$	0.0
$P_{h_{f,G}}(S)$	0.0
$H_{h_{f,G}}(S)$	1.057

Оказалось, что ни в одной из 10 000 выборок множества $\mathcal{S}_{stats}$ не нашлось фиктивной закономерности сильнее, чем в исходной выборке относительно показателя C4. Для оценки ЭМТ в таком случае отлично подходит показатель $H_{h_{f,G}}$, согласно которому скорректированный уровень значимости найденной закономерности составляет ≈ 0.008.

Несмотря на то, что фактические значения функционала $Q(S)$ для фактора C4 и показателей sO2, FHHb ниже, чем таковые для S-100, вероятность получить более высокие значения функционала случайно оказалась ниже. Связано это с особенностью топологии имеющихся данных. Построив аналогичные 1 гистограммы для пары C4–VEGF, удалось обнаружить, что распределение модуля робастного коэффициента корреляции Пирсона сильнее сдвинуто к нулю, чем таковое для пары S-100–VEGF, что и объясняет наблюдаемый эффект:

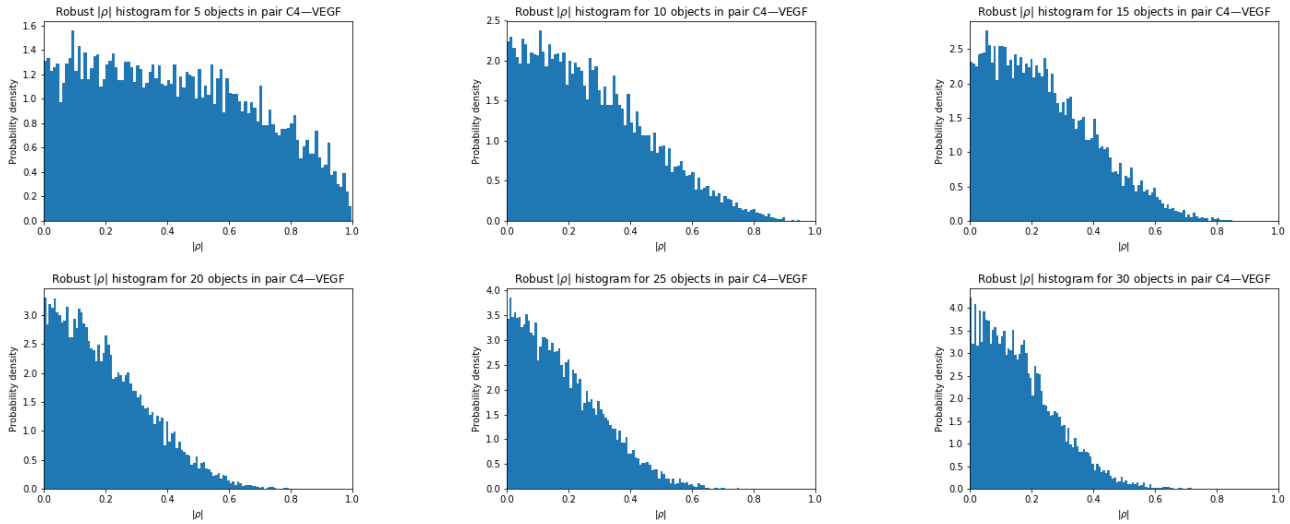


Рис. 4: Гистограммы плотности распределения модуля робастного коэффициента корреляции Пирсона (РККП) (обозначено как $|\rho|$) для выборок разного размера. Каждая гистограмма была получена путём рассчёта РККП для 10 000 выборок фиксированного размера, состоящих из пар случайно выбранных значений показателей C4 и VEGF из исходной выборки данных. На оси абсцисс отложены значения модуля РККП, на оси ординат — значения плотности распределения.

5.1 Дополнительные исследования

Проверяя гипотезы вида «условно-линейной зависимости между целевым признаком (VEGF) и признаком f относительно группы показателей оксиметрии нет», удалось получить весьма значимые результаты для признаков f = S-100 и f = C4. В качестве дополнительного исследования выявленных закономерностей были проверены две гипотезы следующего вида: «условно-линейной зависимости между целевым признаком (VEGF) и признаком f *в совокупности* относительно *полной* группы показателей оксиметрии нет», где под полной группой показателей оксиметрии подразумевается набор признаков pCO2, pO2, sO2, FO2Hb, FCOHb, FMetHb, FHHb, а f принимает обозначенные выше значения. Связь между VEGF и белками S-100 и C4 можно считать доказанной, однако проверка данных гипотез позволяет ответить на вопрос о том, насколько существенную роль в данной связи играет оксиметрия.

Для проверки гипотез использовалась практически та же самая процедура, которая была подробно описана в предыдущих разделах за тем лишь исключением, что вместо перемешивания значений целевого признака VEGF переставлялись значения показателей оксиметрии относительно остальных признаков. В результате проведения перестановочного теста из 100 000 перестановок были получены следующие результаты: для признака S-100 значимость (конкретно — показатель $P_{p_{f,G}}$) оказалась на уровне ≈ 0.0387, для C4 — ≈ 0.0005. Это даёт основание полагать, что в связи VEGF–C4 оксиметрия играет существенно большую роль, чем в связи VEGF–S-100.

Проведя аналогичное исследование, переставляя значения свободного признака f (здесь снова f — либо S-100, либо C4) относительно фиксированных значений VEGF–оксиметрия, было установлено (для обоих исследуемых f), что уровень значимости гипотез «VEGF *в совокупности* с показателями оксиметрии (условно-линейно) не зависит от признака f» оказался менее $2 \cdot 10^{-5}$. Исходя из этого, можно утверждать, что в исследованных закономерностях свободные признаки (S-100 и C4) играют ключевую роль.

Отвержение гипотез всех трёх упомянутых видов для исследованных условно-линейных закономерностей VEGF–S-100–оксиметрия и VEGF–C4–оксиметрия даёт полное основание считать наличие данных связей подтверждённым (во всяком случае, для C4 — точно, так как для всех трёх гипотез уровень значимости не превзошёл $5 \cdot 10^{-4}$, что является весьма сильным результатом, даже учитывая эффект множественного тестирования).

6 Заключение

Для решения задачи подтверждения присутствия условно-линейной зависимости между белками VEGF и S-100 относительно группы показателей оксиметрии был усовершенствован и успешно применён метод Оптимальных Достоверных Разбиений. Дополнительным немаловажным открытием стало нахождение и верификация даже более значимой зависимости подобного вида между белками VEGF и C4.

Итогами данной дипломной работы являются:

1. Основной практический результат: разработаны и протестированы на реальной задаче модификации метода Оптимальных Достоверных Разбиений, позволяющие находить и верифицировать условно-линейные зависимости в данных, а также предложен способ учёта группового эффекта в закономерностях (ранее была возможность исследовать лишь одномерные и двумерные закономерности по отдельности).

2. Усовершенствованный и более теоретически корректный способ поправки уровней значимости найденных закономерностей в связи с эффектом множественного тестирования.

3. Готовый к практическому использованию, отлаженный и вычислительно эффективно реализованный программный продукт, реализующий функционал усовершенствованного метода ОДР.

4. Ряд теоретических результатов, подробное описание которых выносится в Приложение А и В.

Сам метод ОДР и потенциальные сферы его применимости представляют весьма перспективную область для дальнейших исследований. Данная работа является очередным и, будем надеяться, не последним шагом в развитии данного направления.

Список литературы

[1] *Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow and Terence P. Speed.* Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12(2002), 111-139.

[2] *Кодрян М. С.* Статистическая верификация закономерностей в данных при помощи метода оптимальных достоверных разбиений с учётом эффекта множественного тестирования. 2017.

[3] *Сенько О. В.* Перестановочный тест в методе оптимальных разбиений. ЖВМиМФ, 43, 9. 2003. С. 1438–1447. `www.ccas.ru/frc/papers/senko03jvm.pdf`

[4] *Kuznetsova, AV and Kostomarova, IV and Sen'ko, OV.* Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. Springer. Pattern recognition and image analysis, 24, 1, p. 114–123. 2014.

[5] *Морозов А. М.* Разработка методов верификации сложных закономерностей. Кафедра математических методов прогнозирования. 2016.

[6] *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi.* The effects of vascular endothelial growth factor on dendritic cells in esophageal tumor tissue. https://www.ncbi.nlm.nih.gov/pubmed/17210106

[7] *Maya Gulubova, Koni Ivanova, Julian Ananiev, Julieta Gerenova, Aleksandar Zdraveski, Hristo Stoyanov, Tatyana Vlaykova.* VEGF expression, microvessel density and dendritic cell decrease in thyroid cancer. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433839/

[8] *Yunjuan Sun, Kunlin Jin, Lin Xie, Jocelyn Childs, Xiao Ou Mao, Anna Logvinova, and David A Greenberg.* Vegf-induced neuroprotection, neurogenesis, and angiogenesis after focal cerebral ischemia. The Journal of clinical investigation, 111(12):1843–1851, 2003.

[9] *Ingo Marenholz, Claus W Heizmann, and Gunter Fritz.* S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). Biochemical and biophysical research communications, 322(4):1111–1122, 2004.

[10] *Ma L, Chen Y, Jin G, Yang Y, Ga Q, Ge RL.* Vascular Endothelial Growth Factor as a Prognostic Parameter in Subjects with "Plateau Red Face". https://www.ncbi.nlm.nih.gov/pubmed/25919013

[11] *Сенько О. В., Морозов А. М., Кузнецова А. В., Клименко Л. Л.* Оценка эффекта множественного тестирования в методе оптимальных достоверных разбиений. 2016.

[12] *Good, Phillip.* Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media. 2013.

[13] *Grzegorz A. Rempala and Yuhong Yang.* On Permutation Procedures for Strong Control in Multiple Testing with Gene Expression Data. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873102/

[14] *Yoav Benjamini and Daniel Yekutieli.* The control of the false discovery rate in multiple testing under dependency. Annals of statistics, pages 1165–1188, 2001.

[15] *Alexander Y Gordon and Peter Salzman.* Optimality of the holm procedure among general step-down multiple testing procedures. Statistics & probability letters, 78(13): 1878–1884, 2008.

[16] *Olive Jean Dunn.* Multiple comparisons among means. Journal of the American Statistical Association, 56(293):52–64, 1961.

[17] *Sture Holm.* A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, pages 65–70, 1979.

[18] *Peter H Westfall and S Stanley Young.* Resampling-based multiple testing: Examples and methods for p-value adjustment, volume 279. John Wiley & Sons, 1993.

[19] *Dudoit, Sandrine and Shaffer, Juliet Popper and Boldrick, Jennifer C.* Multiple hypothesis testing in microarray experiments. Statistical Science. JSTOR. p. 71–103. 2003.

A О функционалах в методе ОДР

В данном приложении приводятся некоторые теоретические факты о функционалах качества разбиений, использующихся в методе ОДР.

Пусть $F(S,b)$ — некоторый заданный функционал качества разбиения в методе ОДР для данного разбиения b (определяется границами в обобщённом смысле) выборки S, $b_{opt}(F,S) = \arg\max_b F(S,b)$, $F_{opt}(S) = F(S, b_{opt}(F,S))$, $g(x)$ — некоторая монотонно возрастающая функция, $G(S,b) = g(F(S,b))$ — новый функционал. Тогда:

Лемма 1. $F(S,b_1) \leq F(S,b_2) \Leftrightarrow G(S,b_1) \leq G(S,b_2)$

Доказательство. Непосредственно из свойств монотонного преобразования вытекает требуемое утверждение. ∎

Следствие 1. $b_{opt}(F,S) = b_{opt}(G,S)$ ⇒ *можно определить общее для всех эквивалентных с точностью до монотонного преобразования функционалов оптимальное разбиение:* $b_{opt}(S) = b_{opt}(F,S) = b_{opt}(G,S)$.

Лемма 2. $G_{opt}(S) = g(F_{opt}(S))$

Доказательство. $G_{opt}(S) = G(S, b_{opt}(S)) = g(F(S, b_{opt}(S))) = g(F_{opt}(S))$ ∎

Следствие 2. $F_{opt}(S') \leq F_{opt}(S'') \Leftrightarrow G_{opt}(S') \leq G_{opt}(S'')$

Следствие 3. $\mathsf{P}\left(F_{opt}(S_{rand}) \geq F_{opt}(S)\right) = \mathsf{P}\left(G_{opt}(S_{rand}) \geq G_{opt}(S)\right)$, *т. е. p-значения не меняются при монотонном преобразовании функционала.*

Выделим некоторое множество $\mathbf{S}$ выборок. Определим величину $h_{\mathbf{S}}(F,S) = \frac{F_{opt}(S)}{\max_{\tilde{S}\in\mathbf{S}} F_{opt}(\tilde{S})}$.

Лемма 3. $\mathsf{P}\left(h_{\mathbf{S}}(F,S_{rand}) \geq (h_{\mathbf{S}}(F,S)\right) = \mathsf{P}\left(h_{\mathbf{S}}(G,S_{rand}) \geq (h_{\mathbf{S}}(G,S)\right)$, *т. е. вероятность случайного получения лучшего h-значения не меняется при монотонном преобразовании функционала.*

Доказательство. $h_{\mathbf{S}}(F,S') \leq h_{\mathbf{S}}(F,S'') \Leftrightarrow F_{opt}(S') \leq F_{opt}(S'')$ ∎

Таким образом, *монотонные преобразования функционала качества не меняют сути работы метода и интересующие результаты остаются прежними.* К примеру, не имеет смысла подбор коэффициента κ в функционале $\exp\left(-\frac{1-\max(\rho_l^2,\rho_r^2)}{(\rho_l-\rho_r)^2}\frac{\kappa}{\min(m_l,m_r)}\right)$ (данный функционал рассматривался как один из претендентов на место функционала (1)).

В Об оценке ЭМТ в методе ОДР

В данном приложении описывается и подвергается критике процедура учёта эффекта множественного тестирования в методе ОДР, а также предлагается новая процедура, использованная в данной работе.

Для начала стоит подробнее остановиться на p-значениях. Часто p-значения интерпретируют как вероятность наблюдаемой конфигурации при условии истинности нулевой гипотезы H_0. Разберём подробнее данный тезис на примере задач поиска закономерностей в данных.

Итак, положим, рассматриваются «случайные» выборки вида $S = \{(\vec{x}_i, y_i)\}_{i=1}^N$ (в случае ОДР вектор $\vec{x}$ имеет размерность 1 или 2, но в данном случае это не столь важно). «Случайными» данные выборки являются в том смысле, что между признаками $\vec{x}$ и ответом y нет никакой явной закономерности. Выделим совокупность подобных выборок $\mathcal{S} = \{S\}$. Выборки в данной совокупности могут быть в некотором смысле похожими (например, иметь общую конфигурацию признаков, топологию, допустимые значения, связи между нецелевыми признаками и так далее), но все они имеют «случайную» структуру с той точки зрения, которая указана выше.

Основная гипотеза H_0 в задаче ОДР гласит: «связи между нецелевыми признаками и ответом в данных нет, структура выборки — случайна». Фактически это означает, что при выполнении нулевой гипотезы имеющаяся выборка принадлежит множеству $\mathcal{S}$.

Если ввести на основе множества $\mathcal{S}$ вероятностное пространство с элементарными исходами — «случайными» выборками, то можно утверждать, что при истинности нулевой гипотезы имеющаяся выборка была получена в результате некоторого вероятностного эксперимента (с точки зрения общей теории вероятности): то есть в результате некоторого случайного фактора в введённом вероятностном пространстве был выбран элементарный исход — выборка S.

Столь долгое предисловие, возможно, излишне громоздко и несколько запутано, но необходимо для корректности (хотя бы относительной) дальнейших рассуждений.

Итак, имеется вероятностное пространство «случайных» выборок. Введём на нем случайную величину $\xi(S)$ — оптимальное значение некоторого функционала качества, использующегося в методе ОДР — с функцией распределения $F_\xi(x)$. Теперь можно конкретно сформулировать, чем же на самом деле является p-value.

Определим p-значение для данной выборки S так, как это делалось раньше в методе ОДР, но в новых терминах:

$$p(S) = \mathsf{P}(\xi(S) \leq \xi(\tilde{S}) \mid \tilde{S} \in \mathcal{S}) = 1 - \mathsf{P}(\xi(\tilde{S}) < \xi(S) \mid \tilde{S} \in \mathcal{S}) = 1 - F_\xi(\xi(S)).$$

Тогда вероятность события «получить p-значение, не превышающее некоторого уровня значимости α, при условии справедливости нулевой гипотезы (в выборке нет закономерностей, она случайна)» выражается следующим образом:

$$\mathsf{P}(p \leq \alpha \mid H_0) = \mathsf{P}(p(S) \leq \alpha \mid S \in \mathcal{S}) = \mathsf{P}(1 - F_\xi(\xi(S)) \leq \alpha \mid S \in \mathcal{S}) = \alpha.$$

Последнее равенство справедливо в силу известного утверждения о том, что если ξ — случайная величина, а $F_\xi(x)$ — её функция распределения, то случайная величина $F_\xi(\xi)$ является равномерно распределённой на отрезке $[0, 1]$.

Таким образом, мы пришли к замечательному факту: *при условии истинности нулевой гипотезы вероятность получить p-value не более некоторого α равна α* ***при любом функционале качества закономерности***.

Теперь вернёмся к вопросу множественного тестирования в задаче ОДР.

Интересующий нас вопрос ставится следующим образом: «Какова вероятность получить p-value не более некоторого α в выборке, в которой нет никаких закономерностей?» То есть в полной выборке данных проверяется сразу несколько гипотез $H_0^1, \ldots, H_0^m$. Требуется рассчитать следующую (назовём её ЭМТ-) вероятность:

$$\mathsf{P}\left(\bigcup_{i=1}^{m}(p_i \leq \alpha \mid H_0^i)\right). \tag{3}$$

Заметим, что корректная оценка данной вероятности обеспечивает слабый контроль (так как требуется условие справедливости так называемой полной нулевой гипотезы $H_0^C = \bigcap_{i=1}^m H_0^i$, то есть в данном случае полной независимости ответа y от признаков $\vec{x}$) над метрикой FWER, характеризующей вероятность совершить хотя бы одну ошибку первого рода.

Ранее в ОДР способ оценки ЭМТ-вероятности был следующий: на некотором наборе из n случайных выборок вычисляется величина $\bar{\nu}_\alpha = \frac{1}{n}\sum_{i=1}^n \nu_\alpha(S_i)$, где $\nu_\alpha(S) = \frac{\sum_{i=1}^m \mathbb{I}[p_i(S) \leq \alpha]}{m}$ — доля p-значений в выборке S, которые оказались не больше, чем α; затем упомянутая вероятность оценивается как

$$\mathsf{P}\left(\bigcup_{i=1}^{m}(p_i \leq \alpha \mid H_0^i)\right) \approx 1 - (1 - \bar{\nu}_\alpha)^m.$$

На самом деле данная оценка является чрезывачайно грубой, так как опирается на предположение о независимости гипотез $H_0^1, \ldots, H_0^m$, что в свою очередь опирается на независимость наличия закономерности между целевым признаком и отдельными признаками из $\vec{x}$, хотя в реальных данных это, скорее, исключение, чем правило: действительно, например, если один из признаков линейно зависит от другого, то наличие закономерности с одним влечёт наличие таковой и со вторым — они вовсе не независимы. Если всё же предположить, что гипотезы независимы, то:

$$\mathsf{P}\left(\bigcup_{i=1}^{m}(p_i \le \alpha \mid H_0^i)\right) = 1 - \mathsf{P}\left(\bigcap_{i=1}^{m}(p_i > \alpha \mid H_0^i)\right) =$$

$$= 1 - \prod_{i=1}^{m}\left(1 - \mathsf{P}(p_i \le \alpha \mid H_0^i)\right) = 1 - (1-\alpha)^m.$$

При этом упомянутая выше используемая оценка при повышении сложности перестановочного теста стремится к следующей величине:

$$\lim_{n\to\infty} 1 - (1 - \bar{\nu}_\alpha)^m = 1 - (1 - \mathbb{E}\nu_\alpha)^m = \left\{\mathbb{E}\nu_\alpha = \frac{1}{m}\sum_{i=1}^{m}\mathsf{P}(p_i \le \alpha \mid H_0^i) = \alpha\right\} = 1 - (1-\alpha)^m.$$

Таким образом, доказано, что такая оценка ЭМТ-вероятности является грубой в том смысле, что для её справедливости требуется выполнение предположения о независимости проверяемых гипотез.

Куда более корректным являлся бы следующий метод оценки ЭМТ-вероятности (3): снова сэмплируем n случайных выборок; высчитываем долю тех выборок, для которых *существует p-значение,* ***не превышающее*** α; данная доля является Монте-Карло оценкой искомой вероятности (3) и сходится к ней при увеличении сложности теста. Данная оценка не опирается ни на какие предположения о зависимости гипотез и является абсолютно корректной с точки зрения теории вероятности.

Итак:

1. В рамках предыдущего варианта аппроксимации ЭМТ-вероятности (3) получено точное значение предельной оценки: $1 - (1-\alpha)^m$. То есть *отпадает необходимость в подобном варианте перестановочного теста для учёта ЭМТ, так как теоретически найдено предельное значение оценки, к которой сходится тест при увеличении его сложности.* Несмотря на весьма грубое предположение о независимости проверяемых гипотез, которое должно быть выполнено для справедливости данной оценки, её всё же можно использовать при малых значениях α, поскольку она практически совпадает с поправкой Бонферрони (несовместность и независимость для маловероятных событий — близкие понятия): $\alpha \approx 0 \Rightarrow 1 - (1-\alpha)^m \approx m\alpha$.

2. Предложен более корректный вариант оценки ЭМТ-вероятности (3) посредством перестановочного теста.