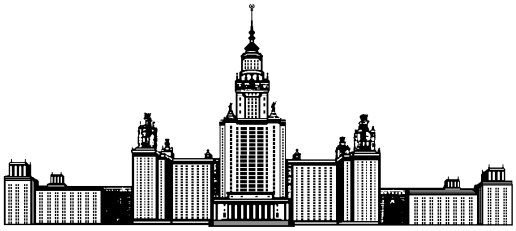


Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Морозов Алексей Михайлович

Разработка методов верификации сложных закономерностей

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д.ф-м.н., в.н.с.
О.В. Сенько

Москва, 2016

Содержание

1 Введение

В современном анализе данных можно выделить два наиболее широких класса задач - прогнозирование целевых переменных по некоторым факторам (т.н. predictive modelling) и составление наиболее полной системы закономерностей между факторами, описывающей некоторое явление. Эти задачи во многом пересекаются друг с другом: для прогнозирования целевых переменных необходимо выделить множество закономерностей, которые в этом помогут, и наоборот - имея систему закономерностей, можно составить прогнозную модель. Однако, эти постановки задачи не являются эквивалентными. В данной работе рассматривается второй подход.

Изначальной предпосылкой данной работы была практическая задача выявления множества логических закономерностей определенного вида в выборке медицинских данных, чтобы в дальнейшем ученые-медики могли использовать эти закономерности в своих исследованиях. Основная задача здесь - определение статистически достоверных закономерностей, то есть таких, которые с низкой вероятностью могли бы появиться в выборке случайно.

В качестве метода поиска логических закономерностей используется метод *оптимальных достоверных разбиений* (Optimal Valid Partition, далее в тексте просто OVP). OVP принимает на вход некоторую выборку данных, состоящую из подпространства факторов из исходной выборки (в данной работе рассматривается 1 или 2 признака) и бинарной целевой переменной. На выходе OVP строит некоторое разбиение подпространства, а также вычисляет некоторое число, являющееся мерой «силы» закономерности. Это число используется в качестве тестовой статистики для проверки следующей гипотезы: пусть зависимости нет, какова вероятность получить настолько сильную закономерность случайно? Для каждого фактора, а также для каждой пары факторов вычисляется эта статистика, а затем из этого множества отбираются значимые закономерности.

Особую роль в таком подходе играет проблема множественного тестирования. Ее можно сформулировать так: при одновременной проверке нескольких гипотез велика вероятность того, что хотя бы одна из них была отвергнута случайно. При этом, в исследованиях обычно требуется гарантия уверенности во всех результатах. Данная работа в основном посвящена построению эффективных методов борьбы с этой проблемой.

Построенный в работе метод прекрасно показал себя на упомянутой практической задаче, однако в исходном варианте он имеет слишком высокую сложность вычислений относительно числа объектов и числа факторов. Поэтому рассмотрены различные способы сокращения сложности вычислений, чтобы сделать метод применимым для выборок различного объема. В частности, исследованы теоретические свойства метода OVP, предложены несколько способов сокращения вычислений. Имеются как строго доказанные методы, так и эвристические, но отлично показывающие себя в экспериментах.

2 Общая постановка задачи и определения

Рассмотрим классическую постановку статистического анализа: есть некоторое множество целевых переменных y, а также набор независимых переменных X. Требуется определить, как независимые переменные влияют на целевые переменные. В зависимости от требований к постановке задачи, можно просто строить некоторую прогнозную модель, а можно искать множество всех закономерностей между совокупностью переменных y и X. Для решения задачи в первой постановке, вообще говоря, не требуется поиск всего множества закономерностей.

Далее в работе приняты следующие обозначения:

- N - число объектов в выборке
- M - число факторов (факторы также часто называют признаками)
- y - некоторое множество целевых переменных. В данной работе рассматривается случай, когда y - это просто бинарная переменная
- $X \in \mathbb{R}^{NxM}$ - выборка
- $f_i(x), i = 1..M$ - множество факторов

Изначальной предпосылкой, как было сказано выше, являлась задача поиска закономерностей для медиков, чтобы они дальше проводили свои эксперименты, опираясь на эти результаты. Таким образом, закономерности должны быть простыми, с ними должно быть удобно работать, и они должны быть легко интерпретируемы. В случае, когда целевая переменная y - категориальная, можно пользоваться аппаратом логических закономерностей. Для построения множества логических закономерностей по данным был использован метод оптимальных достоверных разбиений.

3 Метод оптимальных разбиений

Метод оптимальных разбиений (Optimal Partition, OP) описан в работах [16, 13] . Приведем здесь чуть более общее описание метода из этой работы.

Обозначим за

- S - *split* - семейство допустимых разбиений пространства на k непересекающихся подмножеств
- $s \in S$ - некоторое разбиение
- X_k - объекты выборки, попавшие в k-е подмножество

Введем некоторый функционал $Q(X_k)$, определенный на всевозможных разбиениях пространства X. Важным свойством метода оптимальных разбиений является то, что он допускает введение *любых* функционалов Q. Однако, как правило, функционал Q играет роль меры расстояния между распределением переменной y в множестве X_k и распределением y на всем множестве X.

Введем функционал качества всего разбиения в целом:

- *Интегральный*: $F_I(s, X) = \sum_{i=1}^{k} Q(X_k)$
- *Локальный*: $F_L(s, X) = \max_k [Q(X_k)]$

Оптимальным разбиением называется разбиение

$$s^* = \operatorname*{argmax}_S F(s, X)$$

В дальнейшем под $F(x_1, x_2)$ и $F(x_1)$ будем понимать значения функционалов для этих пар при оптимальном разбиении.

3.1 Примеры функционалов

Как уже было сказано, метод оптимальных разбиений допускает любые функционалы. Однако в задаче анализа данных, являющейся предпосылкой данной работе, целевая переменная y бинарная. Существует множество функционалов Q, разработанных специально для анализа бинарных переменных. Введем еще несколько обозначений:

- N_k - число объектов выборки, попавшее в X_k
- $\nu = \sum_{i=1}^{N} y_i$ - доля объектов класса «1»
- $\nu_k = \sum_{i:x_i \in X_k} y_i$ - доля объектов класса «1» в k-м множестве разбиения

Ниже приведены самые известные функционалы Q для случая бинарной переменной:

- Расстояние между средними: $Q(X_k) = \frac{N_k}{N}(\nu_k - \nu)^2$
- Индекс Джини: $Q(X_k) = -\frac{N_k}{N} 2\nu_k(1 - \nu_k)$
- Энтропия: $Q(X_k) = -\frac{N_k}{N}(\nu_k \log \nu_k + (1 - \nu_k) \log(1 - \nu_k))$

Последние два функционала активно используются при построении решающих деревьев [12]. Заметим, что для достижения наилучшего разбиения энтропию и индекс Джини необходимо минимизировать, поэтому выше они оба взяты со знаком «-».

3.2 Семейства разбиений

Под семейством разбиений мы понимаем множество разбиений с числом элементов, не превышающим некоторое заранее фиксированное число, которые строятся с помощью априори заданного алгоритма. Примеры наиболее простых, а потому наиболее часто используемых семейств разбиений приведены на картинке ниже.

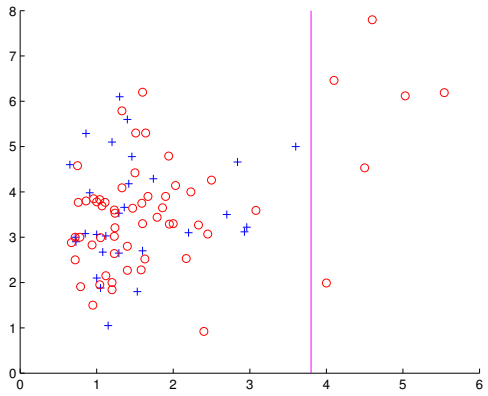


Рис. 1: Семейство 1

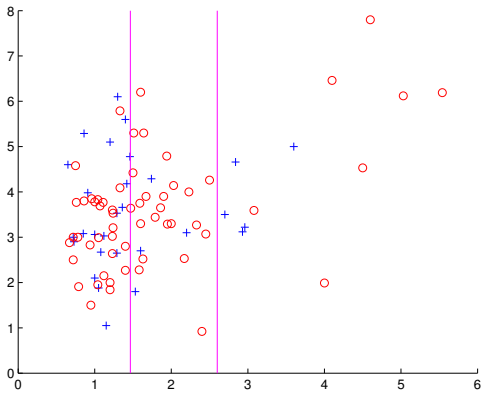


Рис. 2: Семейство 2

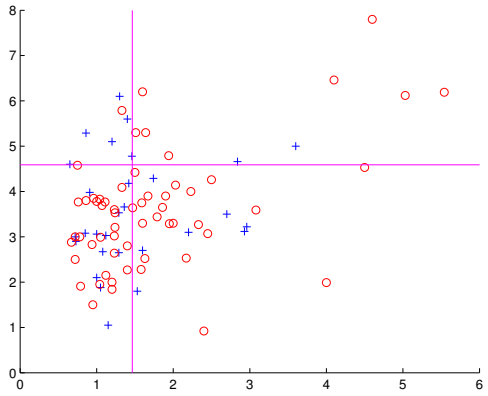


Рис. 3: Семейство 3

- Семейство 1 задает множество всех разбиений выборки не более чем на 2 подмножества, разбиение происходит с помощью граничной точки на оси абсцисс.

- Семейство 2 задает множество всех разбиений выборки не более чем на 3 подмножества, разбиение происходит с помощью двух граничных точек на оси абсцисс.
- Семейство 3 задает множество всех разбиений выборки не более чем на 4 подмножества, разбиение происходит с помощью одной граничной точки на оси абсцисс и одной на оси ординат.

Заметим, что все эти разбиения легко интерпретируемы, следовательно, идеально подходят для решения поставленной задачи. Главным образом в работе изучается 3 семейство разбиений; семейство 2 не рассматривается вообще.

4 Верификация разбиений

Интуитивно, под закономерностью, найденной в данных, понимается некая устойчивая связь факторов, которая должна быть справедлива на всей генеральной совокупности или, иными словами, обладать *обощающей способностью*. Метод оптимальных разбиений позволяет лишь вычислить разбиения, но он не позволяет ответить на вопрос, является оно закономерностью в смысле обобщающей способности. Более того, без каких-либо дополнительных предположений гарантированно ответить на этот вопрос по ограниченной выборке данных невозможно.

Однако, математическая статистика позволяет ответить на другой вопрос: какова вероятность получить такую же или более сильную закономерность при условии, что зависимости между y и X нет? В случае, если такая вероятность мала, то можно утверждать, что данное разбиение действительно обладает обобщающей способностью. Для этого существует аппарат *проверки статистических гипотез* [5]. В терминах статистических гипотез задача формулируется так:

- Задача: проверить нулевую гипотезу H_0 об *отсутствии* зависимости между y и X
- Критерий: если вероятность *случайного* возникновения такой же или более сильной закономерности ниже некоторого уровня α, то гипотеза об отсутствии зависимости отвергается.
- Тестовая статистика T: значение функционала качества $F(s^*, X)$. Это значение отражает «силу» закономерности.
- Математическая формулировка:

$$\mathbb{P}(T \geq T_0) \leq \alpha \rightarrow H_0 \text{ отвергается}$$

Поскольку метод оптимальных разбиений формулируется для любого функционала, то получение аналитических формул распределения невозможно. В таких ситуация естественно использовать перестановочные тесты [6, 8]. Общая схема перестановочного теста такова:

- Вычислить некоторую статистику T_0 на исходных данных
- Сделать некоторое число K перестановок целевой переменной y и вычислить на них значения статистик T_k
- Вычислить *достигаемый уровень значимости* (далее просто p) как долю перестановок, на которых функционал превзошел T_0:$p = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}[T_k > T_0]$

Метод оптимальных разбиений с верификацией закономерностей при помощи перестановочных тестов называется методом оптимальных достоверных разбиений (Optimal Valid Partition, далее просто OVP). Стоит отметить, что, как и метод оптимальных достоверных разбиений, аппарат перестановочных тестов допускает использование *любых* тестовых статистик. Таким образом, метод по-прежнему остается достаточно общим.

4.1 Верификация одномерных закономерностей

Верификация одномерных закономерностей происходит просто перестановками y с тестовой статистикой $F(s, x)$.

4.2 Верификация двумерных закономерностей

Целью исследования пар переменных является выявление значимых *в совокупности* пар, то есть ситуаций, когда оба фактора вносят значимый вклад в закономерность. Ситуации, когда прослеживается явная взаимосвязь одного фактора с целевой переменной, за счет которой парная закономерность тоже получается значимой, должны быть исключены. Поэтому процедура верификации строится следующим образом: для каждого фактора вычисляется свой p - значимость фактора в закономерности. p пары вычисляется как $max(p_x, p_y)$. Таким образом, мы отбираем только те пары, в которых каждый фактор статистически важен.

Чтобы оценить вклад каждого отдельного фактора в закономерность, делается следующая небольшая модификация:

1. Вычисляется разбиение s_0 и значение функционала F_0 для пары.
2. Для каждого фактора:
   - 2.1. Фиксируется разбиение по данному фактору. Это разбиение делит плоскость двух признаков на два подпространства
   - 2.2. Сэмплируется множество перестановок y по следующему правилу: возможны только перестановки меток между объектами, принадлежащими одному подпространству, но нельзя обмениваться метками между объектами из разхных подпространств.
   - 2.3. Вычисляется p

3. p все закономерности вычисляется как $max(p_x, p_y)$.

Ниже представлены примеры правильной и неправильной перестановок (в данном примере фиксируется вертикальное разбиение):

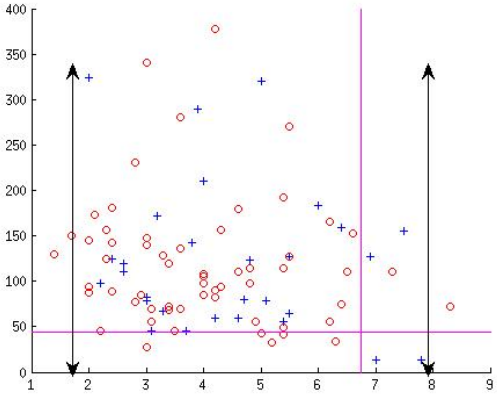


Рис. 4: Правильная перестановка

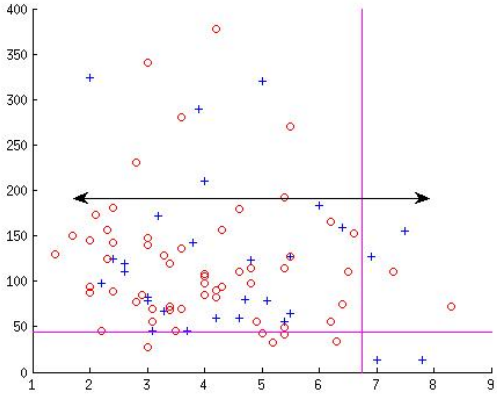


Рис. 5: Неправильная перестановка

5 Проблема множественного тестирования

Проблема множественного тестирования [2, 3, 15] - одна из важнейших проблем прикладной математической статистики. Ее можно объяснить следующим образом: пусть имеется некоторое множество M статистических гипотез, которые мы одновременно проверяем. Ситуация, когда критерий неверно отвергает верную гипотезу, называется ошибкой I рода. Уровнем значимости называется *допустимая* вероятность такой ошибки. Предположим, мы проверяем гипотезы и фиксируем допустимую вероятность ошибки первого рода на уровне $\alpha = 0.05$. Однако, при проверке нескольких гипотез обычно требуется уверенность во всех результатах сразу. Вычислим вероятность появления хотя бы одной ошибки первого рода при M = 20:

$$\mathbb{P}(\#FP \geq 1) = 1 - (1 - 0.05)^M = 1 - (1 - 0.05)^{20} =\approx 0.64.$$

Очевидно, что эта вероятность стремительно возрастает при увеличении числа гипотез.

Для контроля ошибки сразу всех гипотез был разработан целый ряд метрик [3, 2]. В данной работе рассматриваются метрики Familywise Error Rate (FWER), per-Family Error Rate (pFER) и False Discovery Proportion (FDP).

	H_0 верна	H_0 неверна	Σ
H_0 отвергается	V	S	R
H_0 принимается	U	T	m - R
Σ	m_0	m - m_0	m

В введенных выше обозначениях метрики FWER, FDR и FDP соответственно равны:

- $FWER = \mathbb{P}(V >= 1)$
- $FDR = \dfrac{\mathbb{E}(V)}{R}$
- $FDP = \dfrac{V}{R}$

Основное направление работы при исследовании проблемы множественного тестирования - разработка методов, позволяющих контролировать приведенные метрики на заданном уровне α. Рассмотрим некоторые из них.

5.1 Методы, не учитывающие взаимозависимости статистик

Рассмотрим два метода контроля FWER:

Поправка Бонферрони - самый простой метод коррекции FWER [4]. Он основан на том, чтобы просто разделить уровни значимости (или p) на число гипотез M. Теорема Бонферрони гласит, что если гипотезы отвергаются при уровнях значимости $\alpha' = \dfrac{\alpha}{m}$, то FWER $\leq \alpha$. Однако, данная поправка является слишком консервативной: при большом числе гипотез мы практически никогда не будем их отвергать, и суммарная мощность наших критериев будет стремительно падать.

Нисходящие процедуры являются более мягкими по сравнению с поправкой Бонферрони [10]. Более того, показано, что они всегда лучше, чем Бонферрони. Нисходящая процедура множественной проверки гипотез строится следующим образом:

1. Составляется вариационный ряд достигаемых уровней значимости:

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$$

2. По шагам:

   2.1. Если $p_{(1)} \geq \alpha_1$, то принять все нулевые гипотезы и выйти. Иначе отвергнуть гипотезу $H_{(1)}$ и продолжить

   2.2. Если $p_{(2)} \geq \alpha_2$, то принять все нулевые гипотезы, кроме $H_{(1)}$ и выйти. Иначе отвергнуть гипотезу $H_{(2)}$ и продолжить

   2.3. ...

   2.4. Если $p_{(m)} \geq \alpha_m$, то принять все нулевые гипотезы, кроме $H_{(1)}, H_{(2)}, ..., H_{(m-1)}$ и выйти. Иначе отвергнуть гипотезу $H_{(m)}$ и продолжить

В процедуре Холма уровни значимости берутся следующие:

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{\alpha}{m-1}, ..., \alpha_i = \frac{\alpha}{m-i+1}, ..., \alpha_m = \alpha$$

Также как и поправка Бонферрони, метод обеспечивает контроль FWER на уровне α для любых p и любых гипотез.

Поправка Холма равномерно мощнее поправки Бонферрони. Известно, что не делая предположений о зависимостях между статистиками, невозможно построить процедуру мощнее, чем процедура Холма [7].

Перейдем к анализу методов контроля FDR.

Метод Бенджамини-Хохберга - восходящая процедура оценки FDR [9].

1. Составляется вариационный ряд достигаемых уровней значимости:

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$$

2. По шагам:

   2.1. Если $p_{(m)} \geq \alpha_m$, то отвергнуть все нулевые гипотезы и выйти. Иначе принять гипотезу $H_{(m)}$ и продолжить

   2.2. ...

   2.3. Если $p_{(m-1)} \geq \alpha_{m-1}$, то отвергнуть гипотезы $H_{(1)}, H_{(2)}, ..., H_{(m-1)}$ и выйти. Иначе принять гипотезу $H_{(m-1)}$ и продолжить

Модифицированные уровни значимости:

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{2\alpha}{m}, ..., \alpha_i = \frac{i\alpha}{m}, ..., \alpha_m = \alpha$$

Однако, данный метод полагается на одно из двух предположений: либо все тестовые статистики независимы, либо выполнено свойство PRDS [1]. Проверить выполнение свойства PRDS для метода OVP достаточно сложно. Впрочем, в большинстве эмпирических экспериментов проверка этого свойств опускается.

5.2 Перестановочные методы

Для поставленной задачи естественно использовать перестановочный критерий [15, 8, 6], так как множество значений тестовой статистики само по себе определено на всевозможных перестановках метки y. В общем случае перестановочный критерий основан на многократном повторении процедуры поиска достоверных закономерностей на данных с заранее

перемешанными меткам y, где **никаких закономерностей точно нет**. Раз никаких зависимостей нет, то **все** отвергнутые гипотезы будут ошибками I рода. Следовательно, такой способо оценки числа ошибок I рода при уровне значимости α корректен.
В качестве *оценки эффекта множественного тестирования* (далее - оценка МТ) в дальнейшем будет использоваться функция $G(\alpha)$ - эмпирическая оценка ожидаемого числа ошибок первого рода в зависимости от уровня значимости α:

$$G(\alpha) = \mathbb{E}(V|\alpha)$$

Общая процедура

Дано: X, y
Найти: *Распределение вероятности того, что случайно выбранная гипотеза окажется значимой*
for *i = 1: iter_num* **do**
  Случайно перемешать значения целевой переменной y;
  Вычислить p всех гипотез на данных с перемешанными y;
end
Для всех $\alpha = [0; 1]$ вычислить общее по всем итерациям число гипотез, значимых по α.

Algorithm 1: Алгоритм 1: общая процедура оценки эффекта множественного тестирования

5.3 Контроль метрик с помощью перестановочных методов

Опишем, как получить из $G(\alpha)$ уровень значимости α^*, который будет обеспечивать контроль соответствующей метрики.

pFWER Если требуется ограничить ожидаемое число ошибок первого рода на уровне $V^{'}$, то допустимый уровень значимости вычисляется как

$$\alpha^* = \max_{\alpha}\{\alpha : G(\alpha) \leq V^{'}\}$$

FWER Для оценки FWER необходимо знать вероятность ошибки первого рода. Оценку на эту вероятность можно получить, просто разделив $G(\alpha)$ на общее число гипотез m. Вероятность получить хотя бы одну ошибку при данном уровне значимости вычисляется как

$$FWER = 1 - \left(1 - \frac{G(\alpha)}{m}\right)^m$$

Следовательно, для контроля FWER на уровне $\alpha^{'}$ допустимый уровень значимости вычисляется как

$$\alpha^* = \max_{\alpha}\left\{\alpha : 1 - \left(1 - \frac{G(\alpha)}{m}\right)^m < \alpha'\right\}$$

FDP Контроль осуществляется так же, как и для pFWER, с той лишь разницей, что необходимо поделить на число отвергнутых **на исходных данных** гипотез R:

$$\alpha^* = \max_{\alpha} \left\{ \alpha : \frac{G(\alpha)}{R} \leq \alpha' \right\}$$

6 Оценки вычислительной сложности

Предложенный метод дает отличные оценки, однако за них приходится расплачиваться значительным ростом сложности вычислений. Оценим сложность построенной процедуры. Обозначим за

- K_{dist} - число перестановок для оценки распределений функционалов
- K_{MT} - число перестановок для оценки эффекта множественного тестирования

Сложность вычисления тестовой статистики $O(N^2)$. Легко видеть, что сложность всей процедуры равна

$$O(K_{dist} K_{MT} N^2 m)$$

где m, напомним, общее число проверяемых гипотез. В случае полного разведочного анализа данных для всех пар, $m = \frac{M(M-1)}{2}$, что делает процедуру крайне трудоемкой уже при $M \geq 100$.

Очевидно, что такой алгоритм подходит только для выборок ограниченного объема. К тому же, с ростом выборки растет и число перестановок K_{dist}, необходимое для адекватной оценки распределения. Таким образом, за более мягкие оценки эффекта множественного тестирования приходится расплачиваться значительным увеличением объема вычислений.

Цель дальнейшей работы - разработка способов сократить объемы вычислений. В следующей секции описан метод, позволяющий сократить объем вычислений до $O((K_{dist} + K_{MT})N^2 m)$, его обоснование и верификация. Затем предложен ряд методов, позволяющих не вычислять распределения для каждой пары, а вычислить распределения на некоторых синтетических данных и экстраполировать их на остальные гипотезы.

7 Сокращенная процедура оценки MT

Предпосылкой нового метода является следующая идея: если распределение функционала для каждого фактора в паре не зависит от начального разбиения (и, следовательно, от начальной расстановки меток), то на каждой итерации оценки эффекта множественного тестирования распределения статистик будут совпадать. Следовательно, их не нужно вычислять заново на каждой итерации - достаточно вычислить один раз на исходных данных, а при оценке эффекта MT использовать уже посчитанные распределения.

Сокращенная процедура

Оценить распределения тестовой статистики каждой гипотезы на независимых перестановках (для каждой гипотезы свои перестановки).
for $i = 1$: $iter\_num$ **do**
  Случайно перемешать значения целевой переменной y;
  Вычислить сразу все тестовые статистики на перемешанных метках y;
end
Вычислить p.

Algorithm 2: Алгоритм 2: сокращенная процедура оценки эффекта множественного тестирования

Разберем подробнее указанную сокращенную процедуру. При оценке эффекта множественного тестирования изначально требуется случайным образом перемешать метки, а потом повторить всю процедуру верификации. Основная сложность процедуры верификации, очевидно, в оценке распределений статистик для каждого фактора в каждой паре. Если при проверке каждого фактора в закономерности распределение статистики не зависит от начального разбиения, то эти распределения будут одинаковыми для всех итераций оценки эффекта множественного тестирования. Другими словами, можно один раз вычислить распределения и потом их использовать. Сложность в таком случае равна $O(K_{dist}N^2m) + O(K_{MT}N^2m) = O((K_{dist} + K_{MT})N^2m)$. Первое слагаемое - вычисление всех распределений, второе - вычисление всех статистик всех гипотез на итерациях оценки MT. Здесь очень важно отличие процедуры оценки распределений от процедуры оценки эффекта MT. При оценке распределений тестовой статистики для каждой гипотезы нужно сэмплировать независимо от остальных, чтобы в распределениях не было изменений из-за зависимостей между статистиками. При оценке эффекта MT, напротив, необходимо учесть все эти зависимости, поэтому вычисляем все тестовые статистики на одинаковых перестановках y.

Подведем итог представленных рассуждений.

7.1 Требование к функционалу

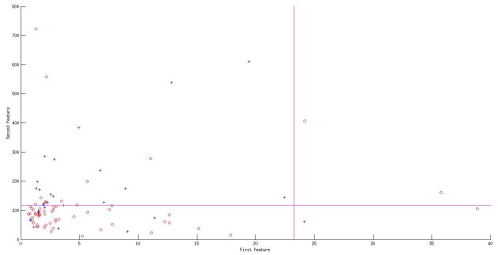
Для использования сокращенной процедуры **необходимо ввести следующее ограничение**: распределение функционала при верификации каждого фактора в паре **не зависит** от начальной расстановки меток и, следовательно, от начального разбиения.

7.2 Анализ старой процедуры верификации

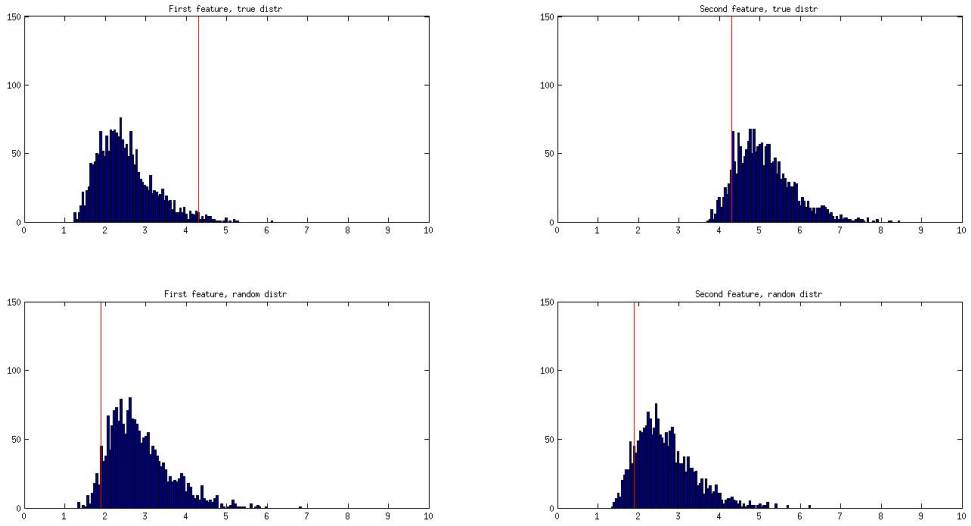
Легко видеть, что исходный функционал не удовлетворяет указанным требованиям. Действительно: при вычислении распределения каждого фактора используются только перестановки специального вида.

Эксперименты показывают, что распределения в разных ситуациях могут быть очень сильно сдвинуты друг относительно друга. Рассмотрим пример:

Разбиение для пары факторов



Распределения для каждого фактора



В true distribution изображены распределения, полученные для исходного изображения выше, а random distributions - для некоторой изначальной случайной перестановки меток. Видно, что распределение для второго фактора сильно сдвинуто в сторону больших значений функционала. Этому есть разумное объяснение: при фиксации разбиения в каждом из подпространств уже получается сильный перекос в сторону одного из классов, это видно по графику. Следовательно, высока вероятность получить хорошее двумерное разбиение, поскольку все перестановки, по которым идет перебор, уже имеют хороший задел для высоких значений функционала.

Первые исследования в работе были направлены на то, чтобы адаптировать данную процедуру под новую постановку и обеспечить выполнение требуемого ограничения. К сожалению, все эти попытки окончились неудачей. Вместо этого возникла другая идея: не пытаться использовать один и тот же функционал для верификации каждого фактора, а ввести **отдельные функционалы для каждого фактора**, которые бы удовлетворяли выдвинутому требованию.

7.3 Описание нового функционала

Введем обозначения:

- $F(x_1, x_2)$ - двумерный функционал
- $F(x_i)$ - одномерный функционал

Введем новый функционал для каждого фактора в паре признаков:

$$F_1(x_1) = F(x_1, x_2) - F(x_1)$$
$$F_2(x_2) = F(x_1, x_2) - F(x_2)$$

Функционал, по которому определяется оптимальное разбиение, остается тем же: $F(x_1, x_2)$. Очевидно, что

$$F(x_1, x_2) - F(x_i) \geq 0 \quad \forall i \tag{1}$$

Действительно, все одномерные разбиения входят в семейство двумерных, если провести одну из линий за областью значений переменных. Максимум по подмножеству не превосходит максимума по всему множеству.

Оба метода верификации на самом деле очень похожи. В обоих методах процедура строится так, чтобы оставлять только те пары, в которых весомый вклад вносят оба фактора. Более того, у нового метода есть еще одна интерпретация. Функционал для верификации каждого фактора можно представить как функционал, зависящий от всей пары, плюс *регуляризатор*, штрафующий ситуации, когда двумерная закономерность является на самом деле отличной одномерной, а второй фактор - просто случайный шум. Интересно то, что в такой ситуации большое p будет как раз у хорошего фактора, а не у плохого. Тем не менее, функционалы такого вида прекрасно справляются с поставленной задачей.

Очевидно, что распределение не зависит от начальной расстановки меток, следовательно, его можно использовать в сокращенной процедуре. Эксперименты показывают, что обычный и новый методы дают практически одинаковые результаты с той лишь разницей, что новый метод считается **на порядки** быстрее.

7.4 Верификация сокращенной процедуры

Чтобы достоверно убедиться в том, что функционал действительно инвариантен, необходимо подтвердить это экспериментом. Был поставлен следующий эксперимент:

1. Случайным образом выбрана 1000 пар
2. На них запущены две версии алгоритма (общая и сокращенная) с большим числом перестановок (чтобы точнее убедиться в сходстве)

3. Были сравнены результаты работы алгоритмов

Результаты верификации показали, что методы абсолютно идентичны и, следовательно, теоретические предположения верны. Подробные результаты верификации описаны в разделе «Эксперименты».

8 Исследование метода OVP

Для начала будет полезно понять, от чего существенно может зависеть само значение функционала OVP

1. Число объектов N.

2. Число повторяющихся значений в факторах. OVP - это максимум по всем возможным разбиениям. Чем меньше повторяющихся значений в каждом факторе, тем больше способов провести разбиение и, следовательно, абсолютное значение функционала может быть выше. Число уникальных значений факторов будем обозначать буквой n.

3. Общее соотношение меток 0 и 1

4. От значений факторов и размещения объектов на плоскости

Распределение значений на перестановках зависит от тех же параметров: соотношения классов в выборке, числа повторяющихся значений в факторах и от размещения объектов на плоскости.

Предположения анализа Будем предполагать, что в рамках одной исследовательской работы метод OVP каждый раз применяется для пар из одной и той же выборки данных и, следовательно, соотношение меток и число объектов N фиксированы для любой пары признаков. Поэтому, будем далее изучать только зависимость от размещения объектов и от числа повторяющихся значений в факторах.

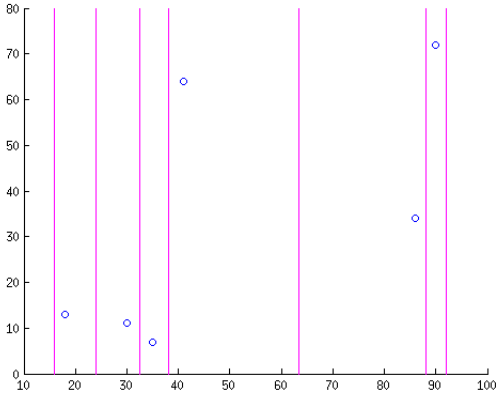
Также предположим, что функционал $Q(X_k)$ на самом деле зависит только от числа объектов и соотношения классов:

$$Q(X_k) = Q(N_k, \nu_k)$$

Применительно к задаче анализа зависимости бинарной переменной практически все разумные функционалы будут удовлетворять этому требованию. Все приведенные выше примеры Q, использованные в данной работе, удовлетворяют ему.

8.1 Анализ модели одномерных разбиений

Допустим, фактор непрерывен ($n = N$). OVP зависит только от множества разбиений объектов:



Следовательно, он инвариантен относительно любых монотонных преобразований значений фактора (монотонные преобразования не меняют взаимное расположение объектов на плоскости). Исследуем, как повторяющиеся значения фактора влияют на значения функционала:

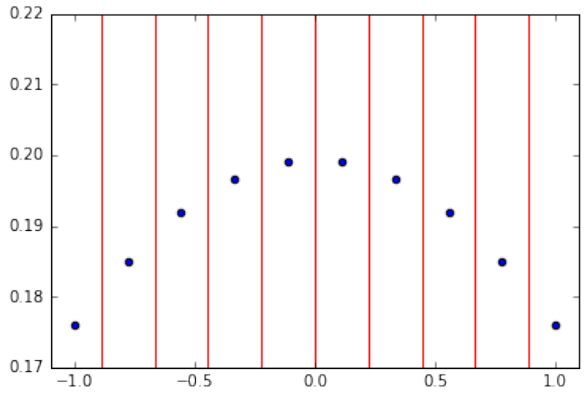


Рис. 6: Все значения уникальны

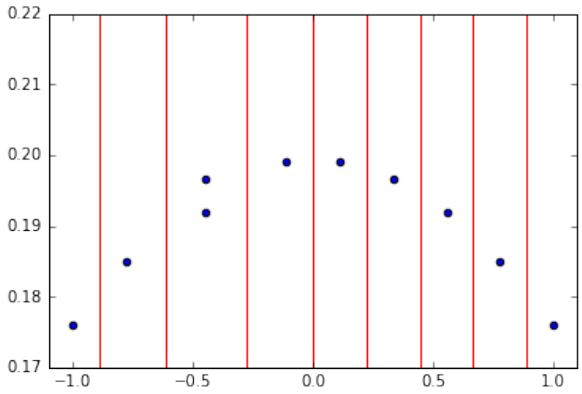


Рис. 7: Есть одно повторяющееся значение

Проанализируем получившуюся ситуацию. При фиксированном числе объектов N имеется ровно N-1 вариант разбить выборку на 2 части. После того, как мы «склеили» два объекта, стало на 1 возможное разбиение меньше. Заметим, что склейка объектов не повлияет на значения функционала при остальных разбиениях: значения функционала зависят только от доли объектов и доли меток. Следовательно, на графике 2 для всех возможных разбиений будут значения функционала *будут совпадать* с значениями функционала для соответствующих разбиений на графике 1.

Полученный результат легко обобщается на случай, когда есть 2 фактора x_1 и x_2 и

множество разбиений x_1 является подмножеством разбиений x_2. В этом случае, если разбиения факторов совпадают, то и значения функционалов совпадают. Если они не совпадают, то $F(x_1) < F(x_2)$. Однако несовпадать они могут лишь в случае, когда разбиение фактора x_2 проходит между объектами, имеющими одинаковое значение в факторе x_1, то есть в факторе x_1 такое разбиение провести нельзя. Из этого следует, что распределение функционала для фактора x_1 можно получить из распределения для x_2: достаточно лишь пересчитать $F(x_1)$ на тех перестановках, на которых разбиение в факторе x_2 не является допустимым разбиением для фактора x_1.

Быстрое вычисление всех одномерных распределений

1. Оценить распределение функционала с большим числом перестановок для некоторого непрерывного искусственного фактора;
for *i = 1:F* **do**
    2.1 Отсортировать значения фактора;
    2.2 Найти повторяющиеся значения и их позиции;
    2.3 Взять перестановки непрерывного фактора;
    2.4 Вычислить заново значения функционалов для тех перестановок, для которых оптимальное разбиение в непрерывном случае проходит между объектами, имеющими одинаковое значения текущего фактора;
end

Algorithm 3: Алгоритм 3: быстрое вычисление всех одномерных распределений

Очевидно, что распределение значений одномерного функционала никак не зависит от начального расположения меток на данных. Следовательно, для одномерного функционала можно применять сокращенную процедуру оценки эффекта MT. Значит, эти два подхода можно скомбинировать и получить очень быстрый и эффективный алгоритм вычисления p вместе с оценкой эффекта множественного тестирования.

Стоит отметить, что в работе [16] предложен алгоритм, вычисляющий точное распределение одномерных разбиений на множестве всех перестановок. Однако сложность этого алгоритма неприемлема уже для $N \geq 40$.

8.2 Анализ модели двумерных разбиений

Проведем анализ для двумерного функционала. Аналогично одномерному функционалу, двумерный зависит только от размещения объектов на плоскости и не зависит от монотонных преобразований обоих факторов. На графике ниже приведена иллюстрация всех возможных разбиений и того, как изменится множество разбиений при «склейке» двух объектов по одному из факторов:

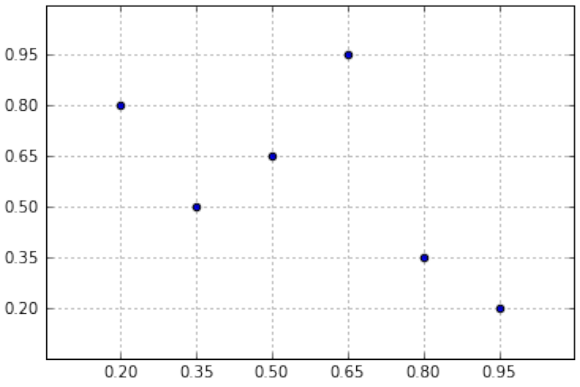


Рис. 8: Размещение объектов

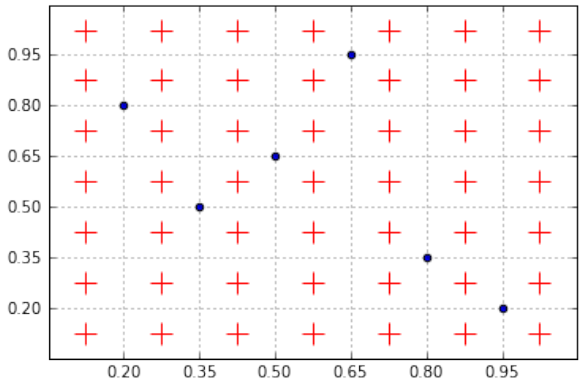


Рис. 9: Множество разбиений

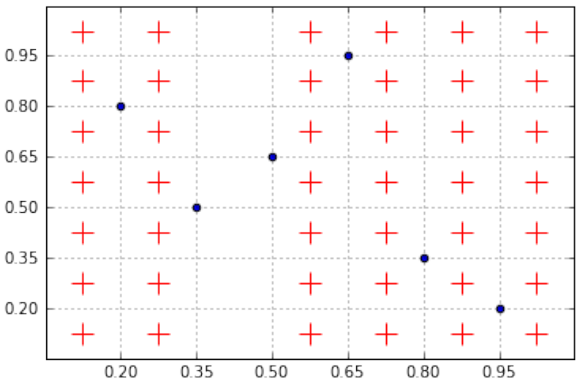


Рис. 10: Множество разбиений при склейке

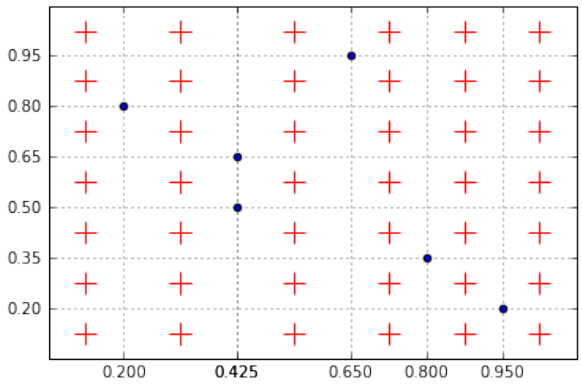
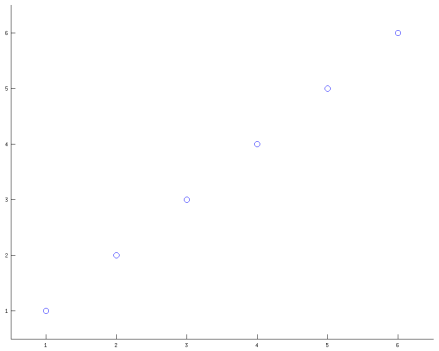
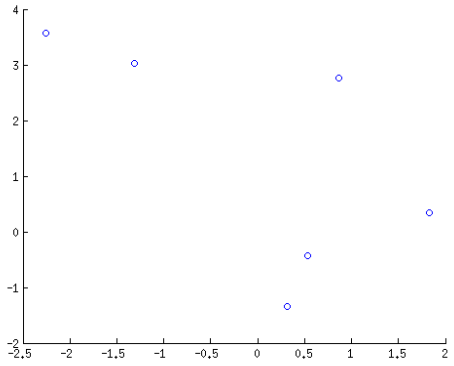


Рис. 11: Множество разбиений при склейке

Отметим одну неприятную особенность двумерного разбиения: даже для двух разных пар факторов с одинаковым числом повторений (для простоты будем считать, что они все непрерывны) распределения функционала на всех перестановках будут немного отличаться. Ниже приведен пример:





Для этих данных были вычислены точные распределения на всех перестановках. Минимальное достигаемое значение функционала на графике слева составляет 0.5, в то время как для графика справа - 0.8333.

8.3 Анализ модели разностного функционала

8.3.1 Оптимизация для большого числа объектов

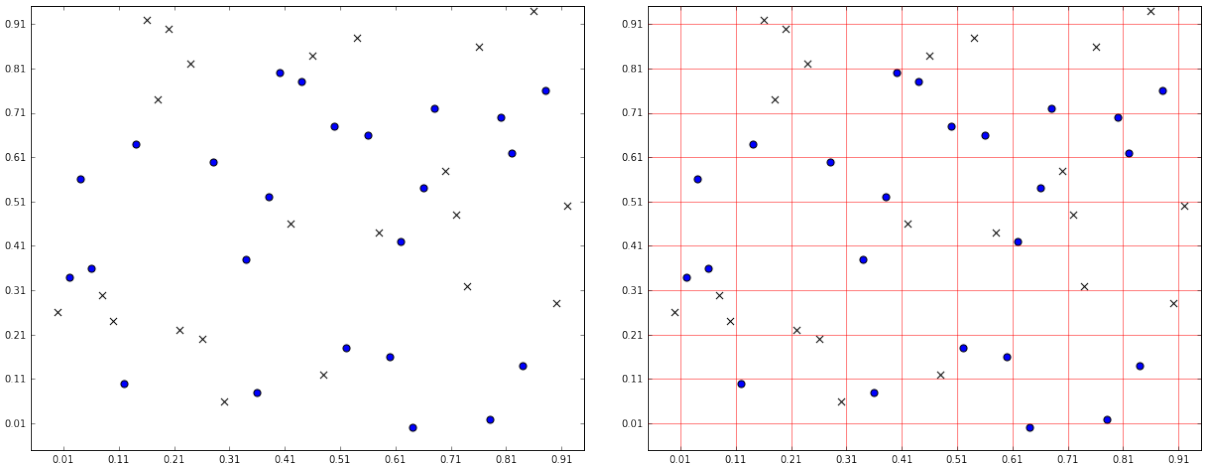
Рассмотрим произвольную пару факторов x_1, x_2. Сложность вычисления максимума среди двумерных разбиений равна $O(n_1 n_2)$. В худшем случае $n_1 = n_2 = N$ и сложность квадратична по числу объектов. Сложностью одномерного функционала можно пренебречь.

Для сокращения вычислений применяется следующая идея:

1. Вычислить быстрым способом $\underline{p}$ - нижние оценки p - для всех пар факторов
2. Исключить из рассмотрения все, у которых $\underline{p} \geq \alpha$
3. Вычислить точные p для оставшихся факторов

Используя этот метод, можно проводить трудоемкие вычисления только для небольшого подмножества факторов.

Быстрый способ предлагается реализовать так: одномерное распределение вычислять точно, а вместо точного вычисления двумерного функционала вычислять его на некотором подмножестве разбиений. Подмножество разбиений предлагается задавать в виде равномерной сетки с некоторым шагом w. Ниже приведены иллюстрации этой идеи:



Если обозначить значение на сетке как $F_{grid}(x_1, x_2)$, то очевидно следующее неравенство:

$$F_{grid}(x_1, x_2) - F(x_1) \leq F(x_1, x_2) - F(x_1)$$

Обозначим за F_0 точное значение разностного функционала на данных. Тогда для любого множества перестановок верно

$$\sum_{i=1}^{K} \mathbb{I}[F_{grid,i}(x_1, x_2) - F_i(x_1) > F_0] \leq \sum_{i=1}^{K} \mathbb{I}[F_i(x_1, x_2) - F_i(x_1) > F_0]$$

Следовательно, $p_{grid} \leq p$, что и требовалось.

Получение нижней оценки p для разностного функционала

1. Вычислить точный максимум $T = F(x_1, x_2) - F(x_1)$;
2. Выбрать число отрезков n_1 и n_2, на которые разобьется множество значений x_1 и x_2 соответственно;
3. Покрыть плоскость равномерной сеткой с выбранными шагами;
4. **for** i = *1:iternum* **do**
   | Вычислить $T_i = F_{grid}(x_1, x_2) - F(x_1)$
   end

Вычислить нижнюю оценку на p $= \dfrac{1}{iternum} \sum_{i=1}^{iternum} \mathbb{I}[T_i > T]$

Algorithm 4: Быстрое получение нижней оценки p разностного функционала

8.3.2 Оптимизация для большого числа факторов

В анализе попарных зависимостей число гипотез равно $\frac{M(M-1)}{2}$. Уже для $M = 500$, что отнюдь не является редкостью в биоинформатике, число пар равно 124750. Основная сложность заключается в том, что для каждой гипотезы нужно вычислять свое распределение тестовой статистики. Однако на самом деле в этом нет необходимости: для пар с приблизительно одинаковым числом повторений распределения тоже будут практически одинаковыми.

Рассмотрим примеры распределений для разностных функционалов OVP с $Q(X_k) = \dfrac{N_k}{N}(\nu_k - \nu)^2$:

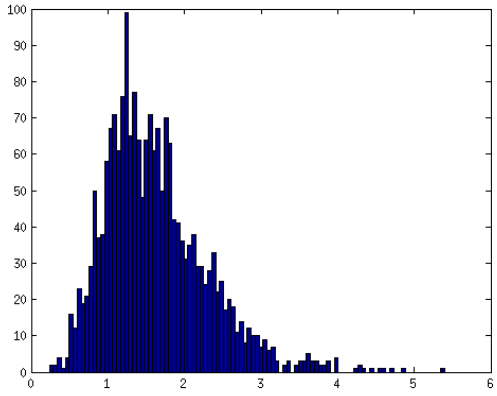


Рис. 12: Интегральный

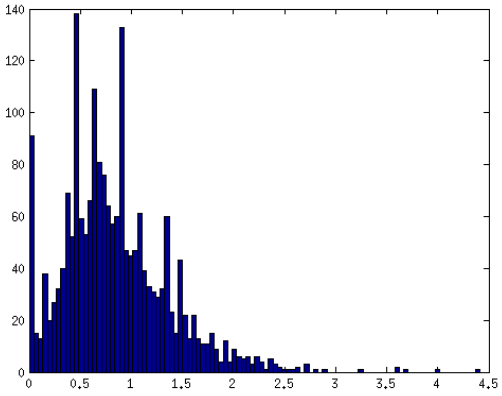


Рис. 13: Локальный

Данные гистограммы получены для одного и того же непрерывного фактора с 2000 перестановками. У локального есть интересный пик в нуле. Все дело в том, что локальный функционал «жадный»: он относительно часто предпочитает некоторые блоки оставлять пустыми в ущерб тому, от которого берется максимум. Интегральный же, напротив, обычно не делает пустых блоков. В результате максимум для двумерного разбиения фактически

равен максимуму для одномерного, так как двумерный сделал один из блоков пустым. Ниже приведены примеры разбиений для интегрального и для локального:

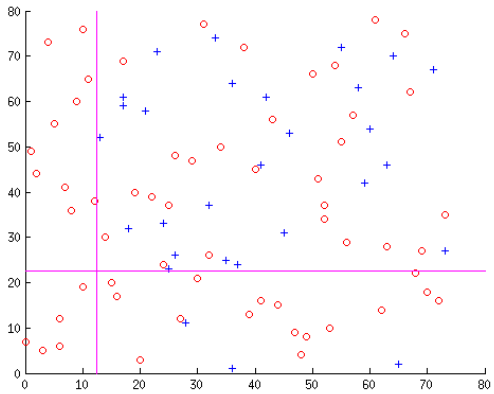


Рис. 14: Интегральный

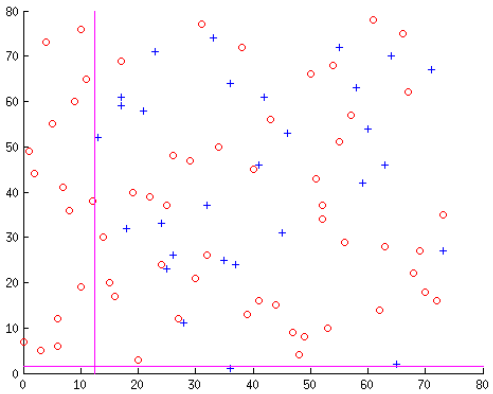


Рис. 15: Локальный

Приведенные выше гистограммы соответствуют двум непрерывным факторам. Ниже приведены графики сглаженных гистограмм:

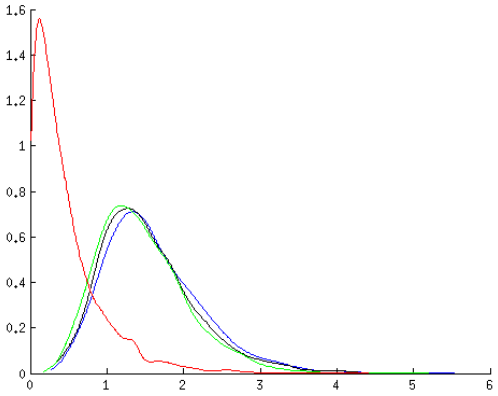


Рис. 16: Интегральный

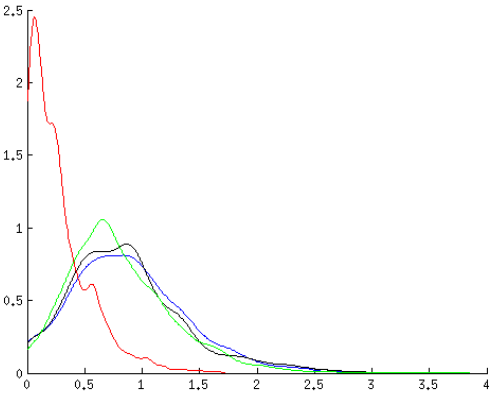


Рис. 17: Локальный

Здесь цветами обозначены плотности для пар, имеющих следующее число уникальных значений:

- Красный - 2
- Зеленый - 32
- Черный - 64
- Синий - 88

Как видно, наблюдается постепенное смещение, которое резко ускоряется при приближении к вырожденным случаям бинарных переменных. Таким образом, видно, что распределения на самом деле очень похожи, за исключением экстремальных случаев.

Теперь более подробно опишем, почему так происходит. Интуитивно понятно, что высокие значения требуют не просто значительного перекоса соотношения классов хотя бы в одном из блоков разбиения, но и большого числа объектов в этом блоке. Таким образом, если рассмотреть две пары непрерывных факторов, то для любого числа объектов найдется разбиение, отсекающее блок определенного размера, поэтому вероятность получить высокое значение сводится, фактически, к вероятности получить высокую концентрацию в блоках определенных размеров. Очевидно, что эта вероятность практически не должна зависеть от размещения объектов на плоскости. То же в некоторой степени верно и для пар факторов, у которых в соответствующих факторах приблизительно одинаковое число повторяющихся значений. Были предприняты попытки строго доказать приведенные интуитивные рассуждения, но все они оказались неудачными. Вместо этого, приведено экспериментальное доказательство этого факта.

Был проведен следующий эксперимент. Были сравнены пары факторов, у которых число уникальных значений в соответствующих факторах отличается не больше, чем на некоторое n. На примере p = 0.05 и 0.01 показано, что распределения порогов для таких уровней значимостей имеет гораздо меньшую дисперсию, чем само распределение функционала. Это означает, что пороги для всех распределений действительно практически одинаковы.

Сначала был проведен следующий эксперимент:

1. Случайным образом сгенерировано 10000 случайный пар с 88 объектами, 29 объектами класса «1» и с непрерывными факторами
2. Для всех пар проведены независимые (на разных перестановках) перестановочные тесты для локального функционала OVP
3. Для всех факторов вычислен порог для p 0.05
4. Изучено распределение порогов

Числа 88 и 29 соответствуют соответствующим параметрам выборки данных, для которой изначально проводилось исследование значимых пар. Ниже представлены два графика: на одном крупным планом изображена гистограмма распределения порогов, на втором - это же гистограмма, наложенная на гистограмму распределения локального функционала:

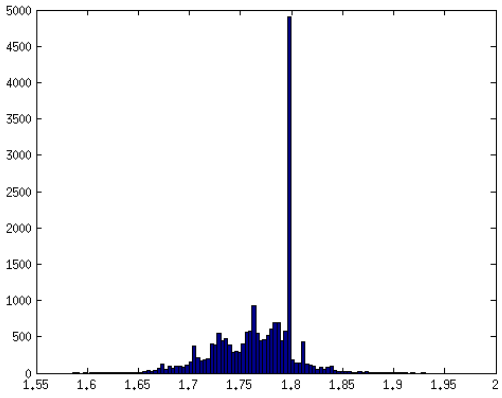


Рис. 18: Распределение порогов

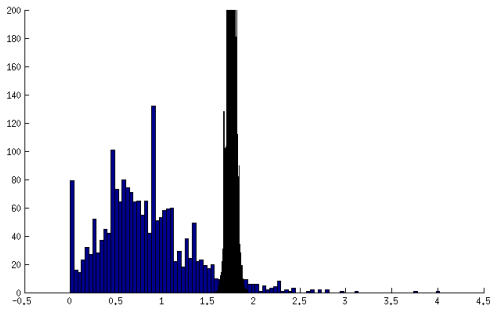


Рис. 19: Распределение порогов на фоне распределения функционала

На графике справа синим цветом изображена гистограмма значений функционала, а черным - гистограмма порогов. Затем задача была усложнена: проведен тот же эксперимент, но с тем отличием, что теперь в факторах могли быть повторения. Максимум повторений - 10.

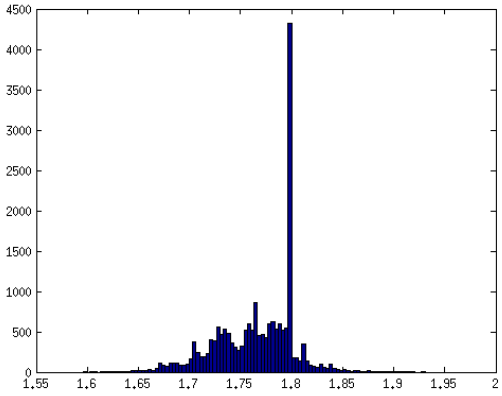


Рис. 20: Распределение порогов

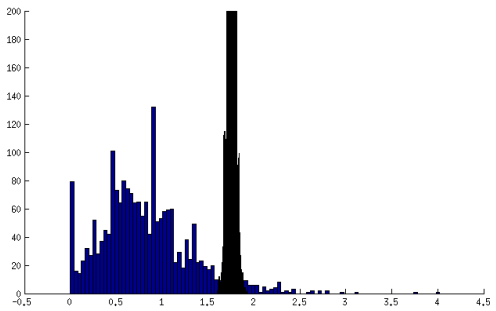


Рис. 21: Распределение порогов на фоне распределения функционала

Эти результаты согласуются с предыдущими: распределения, а следовательно, и распределения порогов, практически совпадают.

Все это вместе дает основания на введение следующей эвристики: будем использовать одно распределение, рассчитанное для некоторой модельной пары, для оценки *всех* распределений, у которых число уникальных значений близко к числам модельной пары. Обозначим за

- $n_{i,model}$ - уник. значения для модельной пары
- $n_{i,real}$ - то же, но для реальных данных

Тогда должно быть выполнено соотношение $n_{i,model} \leq n_{i,real} \leq n_{i,model} + w$, где w-некоторая ширина окна. Поскольку это не строго доказанный факт, а лишь эвристика, то предложенный метод следует применять лишь в случе, когда полный расчет всех распределений действительно невозможен.

Быстрый алгоритм приближенного вычисления p

1. Вычислить точные статистики для всех пар;
2. Выбрать ширину окна w;
3. Вычислить N - число объектов выборки;
4. **for** $i = 0{:}\lceil \frac{N}{w} \rceil$ **do**
   5. **for** $j = i{+}1{:}\lceil \frac{N}{w} \rceil$ **do**
      5.1 $n_1 = \min\{\max\{i * w, 2\}, N\}$;
      5.2 $n_2 = \min\{\max\{j * w, 2\}, N\}$;
      5.3 Сгенерировать случайную двумерную выборку объема N с n_1 и n_2 уникальными значениями факторов;
      5.4 Оценить для нее выборочное распределение функционала;
   end

end

6. Оценить все p для всех пар факторов x_1, x_2, подставляя распределение $n_{i,model} \leq n_{i,real} \leq n_{i,model} + w$;

Algorithm 5: Быстрый алгоритм приближенного вычисления p

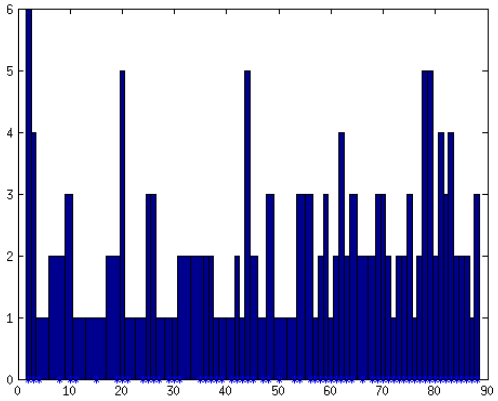
9 Эксперименты

9.1 Описание данных

Помимо модельных данных, для экспериментов использовалась выборка биомедицинских данных, анализ которой являлся предпосылкой данной работы. Параметры выборки:

1. $N = 88$ объектов;
2. $y \in \{0, 1\}$, 29 объектов имеют метку «1»;
3. $M = 143$ фактора;

Биологический и медицинский смысл факторов в работе не рассматривается. Выборка хороша тем, что в ней множество разнородных факторов: бинарные, категориальные, дискретные числовые и непрерывные факторы. Множество мощностей значений факторов довольно богато:



Такое богатое множество факторов делает эту выборку идеальным полигоном не только для отбора закономерностей, но и для тестирования и валидации предложенных в работе методов.

Первоначальная цель медицинского исследования была такова: оценить влияние некоторого фактора S-100 на VEGF в сочетании с другими факторами, иными словами, оценить множество парных взаимодействий, где один фактор фиксирован. Оба фактора часто становятся объектами медицинских исследований [11, 14] VEGF в данном случае выступает целевой переменной, медики из своих соображений бинаризовали его по порогу 750. Однако, помимо этой задачи в работе рассматривается также задача общего разведывательного анализа данных с оценкой множественного тестирования.

9.2 Анализ взаимодействий фактора S-100

Ниже приведен список самых значимых закономерностей в паре с S-100. Обозначим за OVP_{old} - обычный метод верификации, а за OVP_{new} - новый, предложенный в этой ра-

боте. Для OVP_{old} были произведены расчеты для 50 перестановок для оценки эффекта множественного тестирования, для OVP_{new} - 1000.

Таблица 1: Список значимых закономерностей в паре с S-100

Таблица 2: OVP_{old}

Показ.	$p-$знач.
ОЖСС	0.002
S-100	0.002
pCO2	0.013
S-100	0.002
pO2	0.01
S-100	<0.0005
sO2	<0.0005
S-100	<0.0005
FO2Hb	0.025
S-100	0.001
FHHb	0.007
S-100	<0.0005
Ca	<0.0005
S-100	0.007

Таблица 3: OVP_{new}

Показ.	$p-$знач.
ОЖСС	0.0410
S-100	<0.0005
pCO2	0.008
S-100	0.0055
pO2	0.002
S-100	0.0005
sO2	<0.0005
S-100	0.0005
FO2Hb	0.03
S-100	<0.0005
FHHb	0.005
S-100	0.0015
Ca	<0.0005
S-100	0.005

Результаты выше показывают важную особенность: обычный и новый методы верификации дают крайне похожие результаты для самых значимых закономерностей.

Таблица 4: Оценки MT: вероятность случайной значимой пары

	ν, доля случайных значимых пар	
α	OVP_{old}	OVP_{new}
<0.0005	$1.4085 * 10^{-4}$	$4.9296 * 10^{-5}$
0.0005	$1.4085 * 10^{-4}$	$1.3380 * 10^{-4}$
0.001	$1.4085 * 10^{-4}$	$1.3380 * 10^{-4}$
0.002	$2.8169 * 10^{-4}$	$3.3584 * 10^{-4}$
0.003	$2.8169 * 10^{-4}$	$5.7746 * 10^{-4}$
0.007	$8.4507 * 10^{-4}$	0.0011
0.008	$9.8592 * 10^{-4}$	0.0012
0.01	0.0013	0.0017
0.015	0.0018	0.0025

В таблице выше приведены оценки эффекта множественного тестирования, нормированные на число объектов. Иными словами вероятность того, что при данном уровне значимости случайно выбранная значимая закономерность оказалась значимой случайно. Используя эту информацию, вычислим поправленные p, соответствующие различным методам коррекции FWER:

		Поправка		
Пара факторов	p	Бонферрони	Холм	OVP_{old}
S-100/sO2	< 0.0005	0.0705	0.0705	0.0017
S-100/ОЖСС	0.002	0.282	0.28	0.0392
S-100/FHHb	0.007	0.987	0.966	0.1180
S-100/Ca	0.007	0.987	0.966	0.1180
S-100/pO2	0.01	1	1	0.1686

		Поправка		
Пара факторов	p	Бонферрони	Холм	OVP_{new}
S-100/sO2	0.0005	0.0705	0.0705	0.0078
S-100/p02	0.002	0.282	0.28	0.046
S-100/FHHb	0.005	0.987	0.71	0.1017
S-100/Ca	0.005	0.987	0.71	0.1017
S-100/pCO2	0.01	1	1	0.2146

Данная таблица показывает всю мощь перестановочных тестов в методе оценки эффекта множественного тестирования. Легко видеть, что абсолютно все поправки, не учитывающие зависимости между статистиками, не позволяют отвергнуть ни одну гипотезу. Напротив, перестановочный метод дает возможность отвергнуть 2 закономерности на уровне 0.02 и 4 на уровне 0.064. Легко видеть, что всех пар перестановочный тест дает в 10-15 раз более мягкие оценки, чем методы Холма и Бонферрони.

9.3 Общий разведочный анализ выборки

Ниже приведена таблица с числом найденных значимых закономерностей в выборке. Для проверки каждой закономерности использовалось 2000 перестановок.

	$p = 0.01$	$p = 0.05$	$p = 0.1$
OVP, глобальный, старый метод	29	168	393
OVP, локальный, новый метод	30	207	482
OVP, глобальный, новый метод	35	219	491
Энтропия	19	187	368

Две самые значимые закономерности имеют p $<$ 0.0005. Как видно, энтропийный критерий проводит более жесткий отбор пар, чем остальные критерии.

Для оценки всех распределений, как и раньше, использовался перестановочный тест с 2000 перестановок. Таким образом, лучшая оценка на p - это $p < 0.0005$. Приведем лишь скорректированные p для этого уровня значимости:

	p
Коррекция FWER	
Бонферрони	1
Перест. тест	0.4322

К сожалению, даже на уровне значимости 0.0005 не удается отвергнуть ни одной гипотезы. Однако при существенном увеличении числа перестановок вполне вероятно, что отвержение хотя бы одной гипотезы станет возможным.

9.4 Верификация предложенных методов и оценка их качества

В этой секции описаны процедуры верификации и их результаты для предложенных в работе алгоритмов оптимизации.

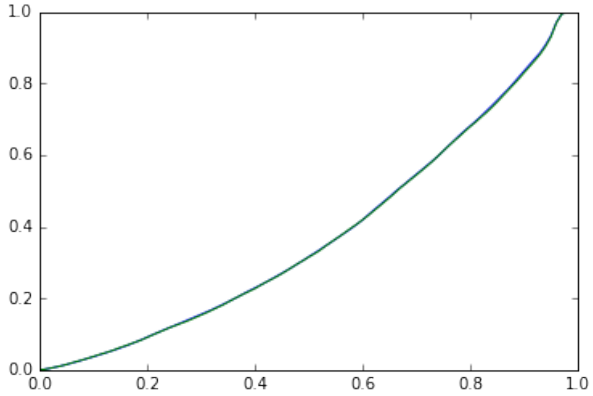
9.4.1 Верификация сокращенного метода множественного тестирования

Данная процедура необходима, чтобы удостовериться в истинности приведенных теоретических рассуждений. Она проводилась на выборке биомедицинских данных. Ход эксперимента таков:

1. Случайным образом выбрана 1000 пар
2. На них запущены две версии алгоритма (общая и сокращенная) с большим числом перестановок.

   - Сокращенная процедура запущена на 1000 случайных перестановках
   - Полная процедура запущена на 50 случайных перестановках

Ниже приведены графики распределения долей значимых закономерностей в зависимости от уровня значимости α.



Как видно, графики полностью совпали. Критерий Колмоговора-Смирнова для функций распределений оценивает p как 1.

Главный результат работы имеет и **теоретическое**, и **доказанное экспериментально** обоснование.

9.4.2 Процедура быстрой оценки одномерных распределений

Данная процедура нужна, во-первых, чтобы оценить правильность теоретических рассуждений, как в случае с введением сокращенной процедуры оценки MT, во-вторых, чтобы оценить прирост в скорости.

1. Выбраны все факторы из исходной выборки
2. На них запущены обе версии алгоритма. Обе с 10000 перестановками

Графики ниже показывают гистограммы разностей p.

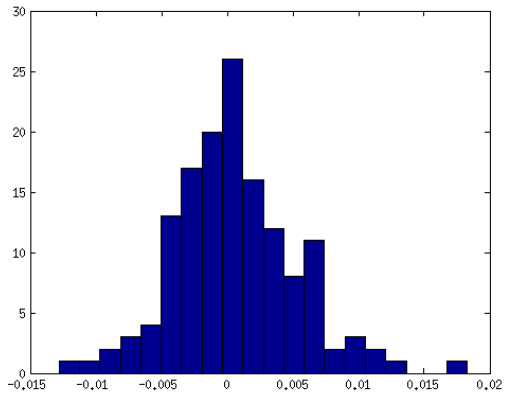


Рис. 22: Разности p для двух запусков полных процедур

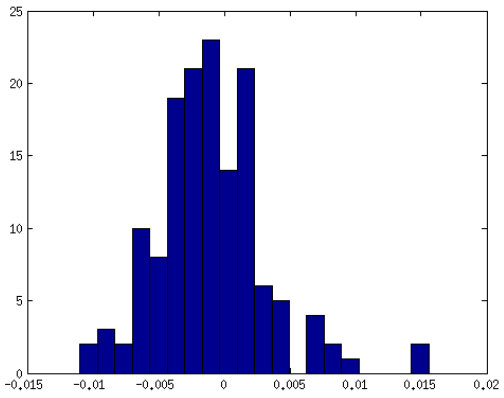


Рис. 23: Разности p для быстрой и полной процедур

- Среднее разности получилось -0.001 для полной и сокращенной процедур. Здесь можно применять статистические тесты, однако все они, за исключением маломощного критерия знаков, отвергнут гипотезу о несмещенности. Все дело в слишком большом числе сгенерированных значений. Плюс ко всему добавляется ошибка сэмплирования

- Оба графика практически совпадают: оба практически несмещены и у обоих совпадает стандартное отклонение. Из этого можно заключить, что метод корректен.

Оценим относительный прирост в ускорении для данного метода. Прирост оценивался для выборки биомедицинских данных. Он вычислялся как отношение сложности обычного метода к сложности ускоренного, обе вычислены точно. Как всегда, обозначим за n_i число уникальных значений фактора i. Результаты такие:

Таблица 5: Относительный прирост для одномерных разбиений

	Прирост
Все факторы	4.8
$n_i \geq 40$	6.3
$n_i \geq 60$	8.38
$n_i \geq 75$	12

9.4.3 Процедура получения нижней оценки p

Здесь ситуация более интересная: помимо проверки правильности теоретических рассуждений, необходимо оценить, насколько заниженными получаются оценки.

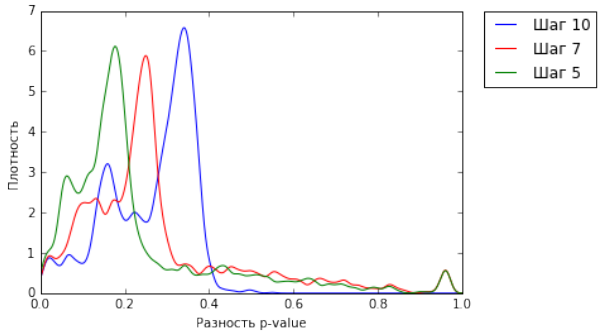
Для верификации процедуры был поставлен эксперимент на модельных данных. Был запущен полный перестановочный тест и модифицированный. Первый эксперимент запущен на модельных данных - все значения функционалов модифицированной процедуры оказались меньше соответствующих значений полной процедуры.

Следующий эксперимент показал точность оценки p при определенном размере окна. Ниже представлены гистограммы разности истинных p и их нижних оценок. Для удобства они были сглажены. Ход эксперимента:

1. Вычислить обычным сокращенным методом распределения тестовых статистик с числом значений ≥ 40.

2. Вычислить сокращенным методом распределения тех же статистик. Были вычислены распределения с шагами сетки 10, 7 и 5.

3. Вычислить разность p обычного и p сокращенного



По сглаженным гистограммам четко видно, что чем меньше шаг сетки, тем лучше аппроксимация. Основная цель, которую должен решать данный подход - фильтрация заведомо плохих закономерностей при больших N и M, когда вычисление всех распределений представляется затруднительным. После работы данного метода можно отсеять все пары, для которых $p \geq \alpha$, и не вычислять для них точные распределения.

Оценим фильтрующее качество метода:

Таблица 6: Доля пар с p < 0.05

	$n_i \geq 40$
Точный метод	0.0434
Сетка с шагом 5	0.123
Сетка с шагом 7	0.152

Учитывая, что вычисление p для сетки ровно в w^2 раз быстрее, чем для полного метода (w - шаг сетки), то можно получить следующую грубую оценку ускорения:

$$0.123 * C + \frac{1}{25} * C = 0.164C$$

где C - максимальная сложность точных вычислений статистик.

9.4.4 Верификация эвристического метода

Данный подход неплохо себя показал в качестве метода вычисления нижней оценки на p.

Достоверная статистическая верификация метода представляется достаточно трудной задачей. В самом деле, во-первых, у метода нет точных теоретических гарантий, поэтому приходится полагаться только на результаты экспериментов. Это значит, что надо исследовать поведение метода при различных N, при различном соотношении классов y, при различной ширине окна w и при различных параметрах факторов. К тому же, одноразовые эксперименты получаются слишком зашумленными, поэтому эксперимент для каждой комбинации параметров следует проводить несколько раз и оценивать средние и доверительные интервалы для отклонений. Такие полномасштабные эксперименты требуют слишком трудоемких вычислений.

Однако для имеющихся данных были проведены некоторые эксперименты. Так, выше был показан пример с распределением порогов для p. При проведении аналогичных экспериментов для признаков с большим числом повторений результаты получаются аналогичные. Однако, выше было показано, что при близком приближении к 2 уникальным значениям распределение начинает резко смещаться к нулю. Поэтому рассуждения о «похожести» распределений верны не всегда.

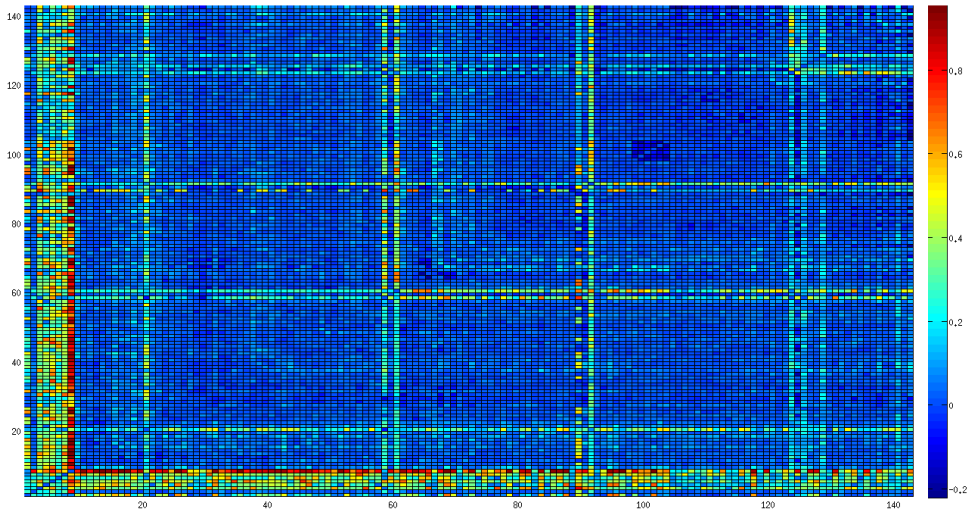
Были проведены следующие эксперименты:

1. Дважды рассчитаны приближенные распределения с шириной окна 2 и 5000 перестановками

2. Один раз рассчитано приближенное распределение с шагом 5 и 5000 перестановками

Анализ p, получаемых при приближенных расчетах с шириной окна 2 показал, что для признаков с $\geq$ 30 уникальными значениями эти распределения дают приблизительно такой же разброс для p, как и точные распределения. Более того, поскольку эффект множественного тестирования позволит достоверно оценивать только закономерности с очень низким уровнем значимости, то необходимо оценивать также совпадение множеств закономерностей с крайне низкими уровнями значимости, так как именно они будут являться результатами экспериментов.

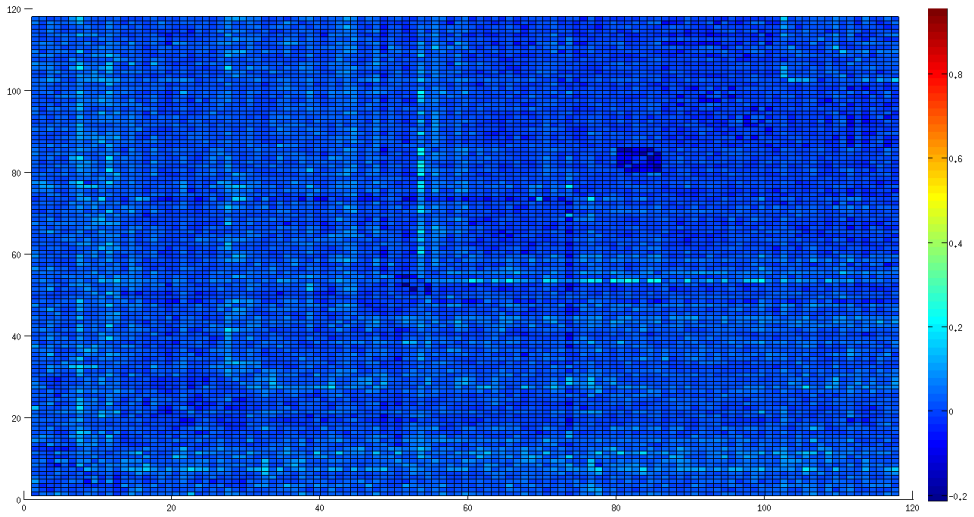
Проанализируем лучшую оценку - приближенное распределение с шириной окна 2. Ниже представлена матрица разностей максимумов p, из p приближенных распределений вычли p точных.



Здесь по осям отложены факторы, на пересечениях - максимум p для пары. Отчетливо

виден характер ошибки: цветовая гамма для каждого фактора приблизительно одинакова, а если мы и ошибаемся, то для всего фактора сразу. Это означает, что на точность метода влияют исключительно параметры фактора.

Дальнейшее изучение показывает, что очень яркие цвета соответствуют в основном факторам с $\leq$ 15 значениями. В основном они сконцентрированы слева и внизу этой матрицы. Если исключить их из выборки, то получается следующая картина:



Ниже представлены гистограммы разности p для разных случаев. В первом отобраны все факторы с $n_i \geq 20$ (около $\frac{1}{4}$ всех значений уникальны), во втором - $n_i \geq 60$
Очевидно, ошибка на графике слева смещена в сторону больших значений. Среднее равно

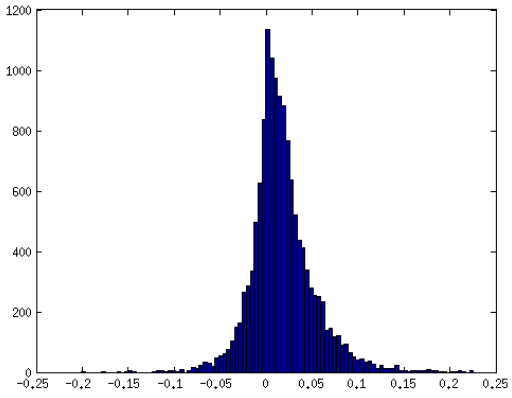


Рис. 24: Разности p, более 20 уник. значений

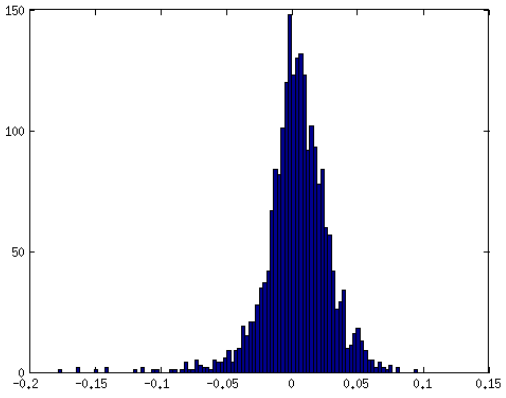


Рис. 25: Разности p, более 60 уник. значений

0.0196. Среднее ошибки на втором графике равно 0.0038. Практически такую же ошибку выдает точное оценивание от запуска к запуску.

Приведенные эксперименты показывают анализ эвристики в целом для всех гипотез.

Однако очевидно, что и для поиска наиболее значимых закономерностей, и для оценки эффекта множественного тестирования по-настоящему важно поведение метода при низких p - всеми остальными случаями можно пренебречь.

В таблице ниже приведено сравнение точного и приближенного методов разведочного анализа данных с использованием локального метода OVP:

Таблица 7: Сравнение точного и приближенного методов при низких p

	Число значимых пар		
p	Точный метод	Приближенный метод	Число совпадений
<0.0005	1	1	1
0.0005	1	2	1
0.001	2	3	2
0.002	4	3	3
0.005	16	8	8
0.01	30	20	20
0.05	205	167	159

Эксперимент показывает, что приближенный метод выделяет **те же** пары, что и точный метод, он лишь является более консервативным.

Оценки ускорения Для рассмотренного выше случая с окном 2 и применением эвристики к факторам с $n_i \geq 40$, получаем, что:

- В точном методе требуется рассчитать $\frac{97*96}{2} = 4645$ распределений для пар с $n_i \geq 40$. В приближенном методе требуется лишь 200 распределений
- При подсчете точных оценок сложности вычислений с учетом сложности вычислений статистик для каждой пары получен относительный прирост почти в 9 раз.

Отметим, что здесь эвристика применялась, по сути, лишь для четверти исходных пар. В случаях, когда в выборке в основном непрерывные пары и факторов много, прирост производительности может быть гораздо больше. К примеру, для $M = 100$, $N = 100$ и таких факторов, что у каждого $n_i \geq 80$, прирост производительности составит 100 раз для окна размера 2 и 24 раза для окна размера 1.

9.5 Выводы

1. Экспериментально показано, что старый и новый методы верификации дают действительно крайне похожие результаты. Фактически, удалось сократить сложность вычислений p вместе с оценкой множественного тестирования до сложности вычислений p без оценки MT

2. Корректность введенного метода доказана не только теоретически, но и эксперимтально

3. Методы Бонферрони и Холма являются крайне консервативными и зачастую не позволяют отвергнуть ни одной гипотезы. Предложенная в работе оценка MT, напротив, дает намного более мягкие оценки, позволяя использовать метод для проверки большого числа гипотез

4. Корректность метод быстрого вычисления одномерных разбиений подтверждена экспериментами. Таким образом, можно дополнительно ускорить проверку одномерных разбиений в несколько раз

5. Метод нижней оценки p дает приемлемое качество фильтрации, однако его применение целесообразно только для очень больших выборок

6. Экспериментально показано, что эвристический метод отбирает те же закономерности, что и точный, однако является чуть более консервативным. Метод следует использовать для разведочного анализа больших выборок данных, когда вычисление всех распределений слишком трудоемко.

7. Показано, что эвристика лучше всего работает для факторов с большим числом уникальных значений n_i. Если в выборке много разнородных факторов, то предлагается следующая стратегия: проводить точный перестановочный тест для всех пар, в которых участвуют факторы с небольшими n_i, а для всех остальных применять эвристику.

10 Заключение

В данной работе предложен новый метод верификации сложных закономерностей, описываемых оптимальными в смысле решения задачи $\operatorname{argmax}_S F(s, X)$ разбиениями для произвольного функционала F, позволяющий на порядки сократить объемы вычислений относительно существующего подхода. Предложенный метод верификации учитывает эффект множественного тестирования. С одной стороны, метод допускает использование практически произвольных статистик. С другой - оценки эффекта множественного тестирования существенно точнее оценок традиционных методов, что позволяет избегать чрезмерной консервативности. Вместе эти процедуры составляют мощное средство для анализа данных на предмет статистически достоверных зависимостей практически любого вида.

Также в работе рассмотрены вопросы сложности вычислений. Основным результатом является разработка метода, позволяющего сократить сложность почти в K_{MT} раз (K_{MT} - число итераций оценки эффекта множественного тестирования). Таким образом, для всех задач стало возможным получать существенно более точные оценки на p-value и эффект множественного тестирования. Однако помимо данных результатов разработаны также дополнительные методики, позволяющие заметно ускорить вычисления. Вместе

введенные методы позволяют применять метод для разведочного анализа больших выборок данных, что раньше не представлялось возможным.

В работе проведена экспериментальная верификация всех оптимизаций вычислений, таким образом, теоретические рассуждения полностью подтвердились в экспериментах.

Наконец, в работе приведен пример решения задачи анализа данных, в которой разработанные методы позволяют получить существенно более качественные результаты, чем традиционные подходы.

10.1 Выносится на защиту

1. Основной результат: разработка метода верификации закономерностей с учетом эффекта множественного тестирования, сокращающего объем вычислений практически до объемов вычислений p-value без оценки множественного тестирования
2. Разработан ряд методов, позволяющих дать существенное дополнительное ускорение метода OVP.
3. Результаты исследования выборки биомедицинских данных: выявлены достоверные с учетом множественного тестирования закономерности, которые не выявлялись с помощью традиционных статистических методов

Список литературы

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

[2] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. CRC Press, 2010.

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

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

[5] Joseph P. Romano E.L. Lehmann. *Testing Statistical Hypotheses*. Springer Series in Statistics. Springer, third edition, 2005.

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

[7] 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.

[8] James J Higgins. Introduction to modern nonparametric statistics. 2003.

[9] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

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

[11] Ingo Marenholz, Claus W Heizmann, and Günter 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.

[12] John Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine learning*, 3(4):319–342, 1989.

[13] Oleg V Senko and Anna V Kuznetsova. The optimal valid partitioning procedures. *Statistics on the Internet http://statjournals. net*, 2006.

[14] 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.

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

[16] Олег В Сенько. Перестановочный тест в методе оптимальных разбиений. *Журнал вычислительной математики и математической физики*, 43(9):1422–1431, 2003.