Комбинаторный анализ эффектов взаимодействия множественных факторов с приложением в генетике

Скурат Евгения Петровна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Алексеева Н.П.
Рецензент: мл. научн. сотр. Ананьевская П.В.

Санкт-Петербург
2013г.

Цель и методы решения

Цель

Решение некоторых актуальных задач, связанных с разработкой конечно-линейного подхода анализа категориальных данных, и его апробация на примере исследования эффектов взаимодействия генетических факторов

Основные обозначения

- Случайный вектор $X = (X_1, \ldots, X_m)^{\mathrm{T}}$ со значениями в $(\mathbb{F}_q, 2^{\mathbb{F}_q})$, заданный на $(\Omega, \mathrm{F}, \mathrm{P})$
- Матрица $\mathbf{A} = \{a_{ij}\}$, $1 \le i \le k$, $1 \le j \le m$, задающая $\tilde{X} = \mathbf{A}X = (X_{\tau_1}, \ldots, X_{\tau_k})^{\mathrm{T}}$: $X_{\tau_i} = a_{i1}X_1 + \cdots + a_{im}X_m$ над $\mathbb{F}_q$

Метод решения

Описание эффектов взаимодействия факторов через линейные комбинации признаков над конечным полем $\mathbb{F}_q$

Дискретная оптимизация в задаче классификации

- Вектор $X = (X_1, \ldots, X_m)^{\mathrm{T}}$ над $(\mathbb{F}_q, 2^{\mathbb{F}_q})$ и строка $\mathbf{A} = \mathbf{A}(1, m)$, $\mathbf{A}X = a_1 X_1 + \cdots + a_m X_m$ над $\mathbb{F}_q$

Поиск наилучшего предсказания конечной дискретной случайной величины Y по $\mathbf{A}X$

Мера отличия двух случайных величин

$$\rho_1(\mathbf{A}X, Y) = \min_{f:\mathbb{F}_q \to \mathbb{F}_q} (1 - \mathrm{P}(\mathbf{A}X = f(Y)))$$

Оптимизационная задача

Поиск точки минимума функции $\sigma(\mathbf{A}) = \rho_1(\mathbf{A}X, Y)$ на множестве строк $\mathbf{A} = \mathbf{A}(1, m)$

Одно из решений опирается на построение алгоритма дискретной оптимизации, основанного на векторной параметризации Грассмана [П. В. Ананьевская, 2013г]

Флаг и согласованность с флагом

Определение

Пусть на $V_m = (\mathbb{F}_q)^m$ *задана последовательность линейных подпространств (**полный флаг** F)*

$$V_0 = \{0\} \subset V_1 = \langle X_1 \rangle \subset \ldots \subset V_m = \langle X_1, \ldots, X_m \rangle$$

такая, что $\cup V_i = V$ *и если* $V_i \subset M \subset V_{i+1}$, *то либо* $V_i = M$, *либо* $V_{i+1} = M$. *Тогда отношение линейного порядка* $\prec$ *называется **согласованным с флагом**, если для всех* $i = 0, 1, \ldots, m-1$ *и* $v \in V_i, w \in V_m \backslash V_i \quad v \prec w$.

Замечание

Выбор флага F *однозначно задает клеточное разбиение **многообразия Грассмана**, определяющего множество всех* k*-мерных подпространств* m*-мерного линейного пространства* [*Ф. Гриффитс, Дж. Харрис, 1982г.*]

Симметричный порядок векторов в пространстве $(\mathbb{F}_q)^m$

- Пространство $V_m = (\mathbb{F}_q)^m$ такое, что $V_m = \langle X_1, \ldots, X_m \rangle$
- Векторы $X_{k_i} = (x_1, \ldots, x_{k-1}, x_{k_i}, 0, \ldots, 0)^{\mathrm{T}} \in V_k \backslash V_{k-1}$, где $x_{k_i} \in \mathbb{F}_q$, $x_{k_i} \neq 0$, $k = 1, \ldots, m$, $i = 1, \ldots q-1$

Определение

Последовательность векторов $\{Y_j\}_{j=0}^{q^m-1}$ *пространства* $V_m = (\mathbb{F}_q)^m$ *обладает свойством* ***симметричного порядка****, если* $Y_0 = \mathbf{0}_m$, $Y_j = sX_{k_i} + Y_t$ *для* $j = sq^{k-1} + t > 0$, *где* $k = 1, \ldots, m$, $i = 1, \ldots, q-1$, $s \in \mathbb{F}_q$, $s \neq 0$, $t = 0, \ldots, q^{k-1} - 1$

Частные случаи

- Лексикографический порядок $X_{k_i} = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$
- Обобщенный порядок Грея $X_{k_i} = (0, \ldots, 0, -1, 1, 0, \ldots, 0)^{\mathrm{T}}$

Теорема о согласованности с флагом

Таблица: Лексикографический порядок над $\mathbb{F}_3$

0	0	0
0	0	1
0	0	2
0	1	0
0	1	1
0	1	2
0	2	0
0	2	1
0	2	2

Таблица: Обобщенный порядок Грея над $\mathbb{F}_3$

0	0	0
0	0	1
0	0	2
0	1	2
0	1	0
0	1	1
0	2	1
0	2	2
0	2	0

Теорема

Симметричный порядок согласован с полным флагом F на пространстве $V_m = (\mathbb{F}_q)^m$.

Вероятность случайной классификации

- Случайные вектор $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ и матрица $\mathbf{X} = \mathbf{X}_{n,m} = (X_1, \ldots, X_m)$ над $\mathbb{F}_q$, заданные на $(\Omega, \mathrm{F}, \mathrm{P})$; X_i независимы и одинаково распределены
- Линейное преобразование $X_\tau = a_1 X_1 + \cdots + a_m X_m$ над $\mathbb{F}_q$

Функция, равная количеству ошибок классификации

- $\rho_1(X_\tau, Y) = \min\limits_{f:\mathbb{F}_q \to \mathbb{F}_q} (1 - \mathrm{P}(X_\tau = f(Y)))$
- $\rho(\mathbf{X}, Y) = \min\limits_{X_\tau \in \mathfrak{L}(\mathbf{X})} \rho_1(X_\tau, Y)$, где $\mathfrak{L}(\mathbf{X}) = \langle X_1, \ldots, X_m \rangle$

  $F(t) = \mathrm{P}(\rho(\mathbf{X}, Y) < t)$ — вероятность случайной классификации

Известны асимптотическая оценка $F(t)$ [Н. П. Алексеева, 2009г.] и верхняя оценка [П. В. Ананьевская, 2013 г.]

Проблема существования точной оценки $F(t)$

Точная оценка вероятности случайной классификации

Задача поиска точной оценки:
Вычисление количества невырожденных матриц $\mathbf{X}_{n,m}$ с весом $L = 1, \ldots, M$ линейной оболочки $\mathfrak{L}(\mathbf{X})$ для нулевого вектора классификации $Y = \mathbf{0}_n$, где $L = \min\limits_{i=1,\ldots,m} l(X_i) = \min\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{n} x_j$

Теорема

Число невырожденных матриц $\mathbf{X}_{n,m}$, порождающих линейную оболочку с весом $L = 1$, вычисляется по формуле:

$$\mathbf{X}_{n,m} = \mathbf{X}_{n-1,m} + \sum_{t=0}^{m-1} C_m^t \cdot ((\mathbf{X}_{n-1,m-t} \cdot (\mathbf{V}_{n-1,t} + t \cdot \mathbf{V}_{n-1,t-1})) +$$
$$+(\mathbf{V}_{n-1,m-t} - \mathbf{X}_{n-1,m-t}) \cdot ((m-t) \cdot t \cdot \mathbf{V}_{n-1,t-1} + t \cdot \mathbf{V}_{n-1,t-1})) +$$
$$+(n-1) \cdot 2^{n-2} \cdot 2 + (\mathbf{V}_{n-1,m-1} - (n-1)) \cdot m,$$

где $\mathbf{V}_{n,m} = \prod_{j=0}^{m-1}(2^n - 2^j)$ — общее число невырожденных матриц.
Тогда точная оценка $F(t) = \mathsf{P}(\rho(\mathbf{X}, Y) = 1)$ имеет вид $\frac{\mathbf{X}_{n,m}}{\mathbf{V}_{n,m}}$.

Апробация разработанных методов. Описание данных

Исследовательские центры: НИИ фармакологии им. А.В. Вальдмана СПбГМУ им. акад. И.П. Павлова и Ленинградский областной наркологический диспансер.

Профилактика рецидива опийной наркомании.

- **Индивиды** — больные героиновой зависимостью ($n = 245$), проходившие курс (26 недель) психотерапии в сочетании с рандомизированным исследованием эффективности налтрексона.
- **Переменные** — гены опиатных рецепторов ($m = 15$), отвечающие за когнитивную функцию, моторику и энергетику.
- **Ковариата** — способы терапии (двойное плацебо, пероральный налтрексон, продетоксон).
- **Итоговые характеристики**
  - количество положительных тестов на опиаты;
  - длительность удержания в программе;
  - отсутствие рецидива.

Проблема анализа эффектов взаимодействия

Проявление совокупного воздействия двух и более переменных не в виде суммы отдельных факторов.

Таблица: Средние количество $(+)$ тестов на героин в сочетаниях генов A (мигрени, беспокойства) и B (никотиновая зависимость).

A	0	0	1	1
B	0	1	0	1
$A+B \pmod 2$	0	1	1	0
среднее	5.0	5.4	6.8	1.0

В явном виде эффекты взаимодействия могут быть выражены как конечно-линейные комбинации над $\mathbb{F}_q$, которые для удобства работы с приложениями названы **симптомами** [Н. П. Алексеева, 2008г.]

Задача и структура решения

Задача

Выявление совокупности генетических факторов, значимо влияющих на тяжесть наркотической зависимости

- Выбор признаков индикатора рецессивности генотипов
- Построение линейных комбинаций признаков над $\mathbb{F}_2$ (симптомов)
  - с ограничением на ранг
  - без ограничения на ранг с применением алгоритма дискретной оптимизации в случае обобщенного порядка Грея
- Исследование влияния симптомов на результат лечения в качестве фактора в статистических критериях
- В задаче классификации проверка случайности относительно итогового фактора безрецидивности

Алгоритм отбора симптомов с ограничением на ранг

- Начальные параметры
  - уровень значимости α
  - информативность симптома M
  - предельный ранг k
- Последовательный перебор симптомов X_τ, $|\tau| = 1, \ldots, k$
- Применение статистического критерия $p = p(X_\tau)$, где симптом выступает в качестве фактора
  - в дисперсионном анализе
  - в анализе данных типа времени жизни
  - в информационной статистике
- Включение значимых симптомов X_τ
  - $p(X_\tau) < \alpha$
  - $H(X_\tau) > M$, где $H(X_\tau) = -\sum\limits_{i=1}^{q} p_i \log_2 p_i$
- Исключение симптомов X_τ, не вносящих дополнительной информации: для $\delta > 0, \epsilon > 0 \quad |\tau| > |\tau_0|$,

$$H(X_{\tau \setminus \tau_0}) < \delta \text{ и } H(X_\tau) - H(X_{\tau_0}) < \epsilon$$

Двухфакторный дисперсионный анализ

Модель с фиксированными эффектами факторов A и B

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \xi_{ijk}$$

y_{ijk} — количество положительных тестов на героин
α_i — дифференциальный эффект фактора A симптома X_τ, $|\tau| < 4$
β_j — дифференциальный эффект фактора B терапии
$(\alpha\beta)_{ij}$ — эффект взаимодействия A и B
ξ_{ijk} — ошибки независимые, $N(0, \sigma^2)$

$X_1, \dots X_m$, $m = 15$ — факторы рецессивности генотипов

$H_0 : \alpha_i = 0$

τ	p	$H(X_\tau)$
(15)	0.043	0.068
$(8, 13)$	0.018	0.196
$(1, 5, 12)$	0.019	0.261

$H_0 : (\alpha\beta)_{ij} = 0$

τ	p	$H(X_\tau)$
(9)	0.021	0.114
$(7, 13)$	0.029	0.135
$(1, 5, 12)$	0.049	0.261

Положительный эффект психотерапии без налтрексона при парном сочетании генов: никот-вая зав-ть (1) , депрессия (5), алког-ая зав-ть (12)

Анализ данных типа времени жизни

Критерий Гехана-Вилкоксона о равенстве медиан продолжительности участия в программе при разной терапии

- Ковариата — симптом X_τ, $|\tau| < 4$
- Правое цензурирование, индикатор — результат выполнения программы
- Интервальное цензурирование, $[t_1, t_2]$, t_1 — точка последнего наблюдения; $t_2 = t_1 + 1$

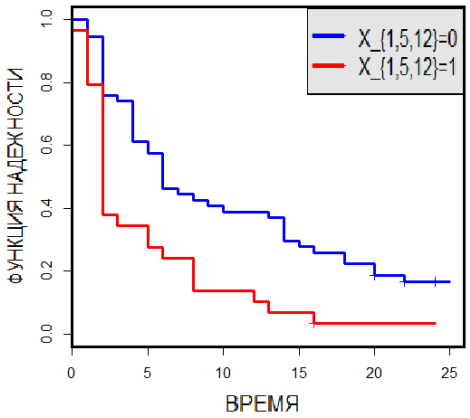

Рис.: Значимое влияние $X_{1,5,12}$ на дожитие (плацебо, p=0.0006)

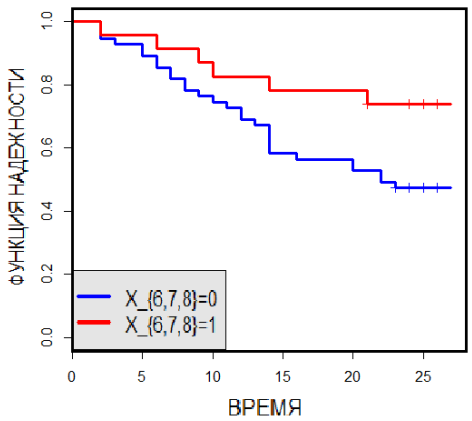

Рис.: Положительный эффект $X_{6,7,8}$ при продетоксоне, p=0.044

X_6 – гиперактивность, X_7 – болезнь Паркинсона, X_8 – шизофрения

Критерий Пирсона независимости категориальных признаков на основе таблиц сопряженности

- Значимое влияние факторов энергетики $X_{1,5,12}$ (p=0.047) и когнитивности $X_{6,7,8}$ (p=0.047) на рецидив
- Количество ошибок прогнозирования рецидива по значимым симптомам X_τ, $|\tau| > k$ и верхние оценки вероятности случайной классификации

X_τ	ошибки	случайность	p	$H(X_\tau)$
$X_{(1,5,12),(7,8),(3,11)}$	93	0.00502	0.031	0.275
$X_{(1,5,12),(7),(3,11)}$	90	0.00049	0.028	0.263
$X_{(1,5,12),(6,7,8),(3,11)}$	101	0.15899	0.043	0.289

X_3 – импульсивность, X_{11} – дискинезия

Взаимодействие генов и эффект лечения

$Y = X_{1,5,12} \oplus X_{6,7,8} \oplus X_{3,11}$		$Y=0$				$Y=1$			
$X_{3,11}$	моторика	0	0	1	1	0	0	1	1
$X_{6,7,8}$	когнитивность	0	1	0	1	0	1	0	1
$X_{1,5,12}$	энергетика	0	1	1	0	1	0	0	1
число инд-дов	245	128	8	11	2	50	40	4	2

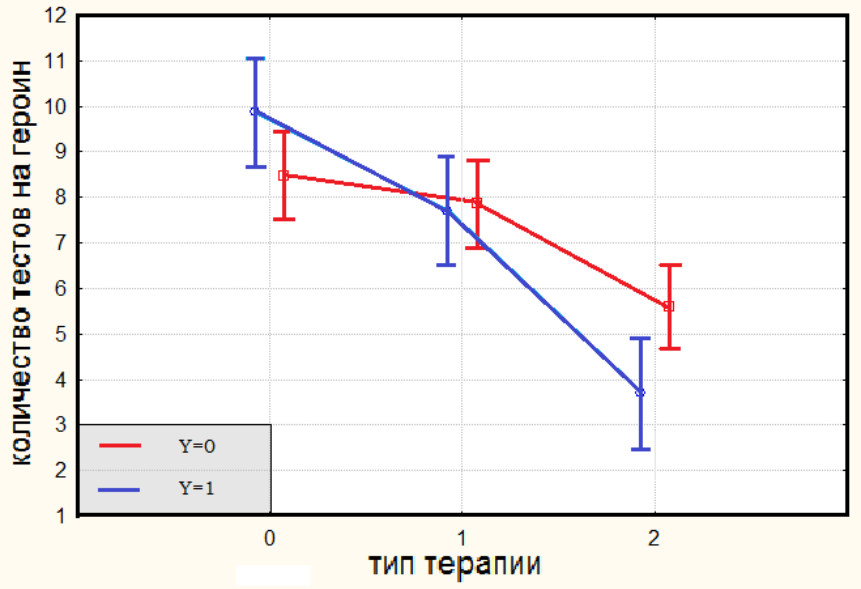

Рис.: Зависимость от продетоксона эффекта лечения при одной генетической особенности, p=0.011.

Основные результаты

- Проведение комбинаторного анализа эффектов взаимодействия множественных факторов на примере данных о программе лечения героиновой наркомании
- Реализация программы разработанного математического метода исследования категориальных данных в статистическом пакете R
- Определение симметричного порядка и обобщение теоремы о согласованности с флагом для введенного порядка
- Доказательство формулы точной оценки вероятности случайной классификации в частном случае