Редукция размерности категориальных данных на основе точного критерия Фишера

Куликов Даниил Владимирович, гр. 422

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Н.П. Алексеева
Рецензент: к.т.н., ст. науч. сотр. Л.А. Белякова

Санкт-Петербург

Постановка задачи

Исследование и реализация методов определения наиболее информативных признаков.

- Изучение точного критерия Фишера для таблиц размерности $r \times c$.
- Программа для вычисления значения точного критерия Фишера.
- Изучение редукции размерностей на основе использования грассманиана.
- Программа поиска наиболее информативного подпространства.
- Использование рекуррентности для параметризации грассманиана.
- Анализ результатов на реальных данных.

Точный критерий Фишера

Таблица сопряженности $x = \{x_{ij}\}_{i=1,j=1}^{r,c}$ размерности $r \times c$.

- Множество элементарных исходов Ω — таблицы размерности $r \times c$ с маргинальными суммами таблицы x.

$$m_i = \sum_{j=1}^{c} x_{ij}, n_j = \sum_{i=1}^{r} x_{ij}, N = \sum_{i=1}^{r} m_i$$

Нулевая гипотеза: $x_{ij} = m_i \times n_j$ для любых (i, j).

- Степень отклонения $FI(x)$ от нулевой гипотезы всех таблиц $x \in \Omega$.

$$FI(x) = -2\log(\gamma P(x))$$

$$\gamma = (2\pi)^{(r-1)(c-1)/2} N^{-(rc-1)/2} \prod_{i=1}^{r} (m_i)^{(c-1)/2} \prod_{j=1}^{c} (n_j)^{(r-1)/2}.$$

Точный критерий Фишера

- Точное p-значение критерия Фишера

$$p = \sum_{FI(y) \geq FI(x)} P(y)$$

$$y \in \Omega \text{ и } P(y) = \frac{\prod_{i=1}^{r} m_i! \prod_{j=1}^{c} n_j!}{N! \prod_{j=1}^{c} \prod_{i=1}^{r} y_{ij}!}.$$

Проблема: множество Ω растет экспоненциально с ростом размерности.

Решение: алгоритмы перечисления таблиц сопряженности: Агрести А. (1990), Мехта С.Р., Патель Н.Р. (1989), Вербик А. (1985), Пагано М., Хальворсен К.Т. (1981)

Программа и сравнение результатов

Данные: (ПСПбГМУ им. академика И. П. Павлова) о рецидивах заболевания щитовидной железы.

- НИС — натрий-йодный симпортер (определяет возможность захвата йода клетками мембраны).
- T-стадия — одна из четырёх стадий размера опухоли.

Таблица : Собственное сравнение НИС и T-стадии

	Fisher	Fisher Monte-Carlo	Chi-square	Likelihood ratio
p-value	0.749	0.745	0.825	0.926

Таблица : Сравнение НИС и T-стадии в SPSS

	Fisher	Chi-square	Likelihood ratio
Точная значимость	0.749	0.825	0.926

Вычисление информативных признаков с использованием грассманиана

Определение:

Пусть $X_1, \ldots, X_m$ — дискр.сл.вел. со значениями над F_q, тогда линейная комбинация $\sum_{i=1}^{m} a_i X_i \pmod q$, где $a_i \in F_q$ — симптом.

Пусть $X_{\tau_0}, \ldots, X_{\tau_k}$ — линейно независимые симптомы, множество $q^{k+1} - 1$ всех линейных комбинаций симптомов над F_q вида $\{\sum_{j=0}^{k} b_j X_{\tau_j} \pmod q\}$ — синдром порядка k.

Применение: значимость взаимосвязи синдрома с итоговой характеристикой с помощью точного критерия Фишера.

Определение:

Всевозможные k-мерные подпространства пространства $V_m = (\mathbb{F}_q)^m$ образуют грассманиан $Gr_q(k, m)$, точкой которого является одно k-мерное подпространство.

Алгоритм

Задача: алгоритм перечисления точек грассманиана.

Теорема о векторной параметризации (Ананьевская П.В., 2013):

Существует биекция между k-мерными подпространствами V_m и симптомами $(X_{\tau_1}, \ldots, X_{\tau_k}) \in V_m$, заданная матрицей вида:

$$
\begin{array}{c}
 \\
X_{\tau_1} \\
\vdots \\
X_{\tau_i} \\
\vdots \\
X_{\tau_k}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc}
X_1 & \ldots & & & & \ldots & & X_j & & \ldots & & & X_m
\end{array} \\
\begin{pmatrix}
* & \ldots & * & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
* & \ldots & * & 0 & * & \ldots & * & 1 & 0 & \ldots & 0 & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
* & \ldots & * & 0 & * & \ldots & * & 0 & * & \ldots & * & 1 & 0
\end{pmatrix}
\end{array}
$$

Решение: использование алгоритма быстрого перечисления точек грассманиана, основанного на его клеточном разбиении.

Альтернативный алгоритм

Определение:

Клеткой Шуберта называется $S_I \subset Gr_q(k, m)$, состоящее из всевозможных подпространств V_I типа I — вектора длины k из строго возрастающих номеров $1 \leq i_1 < \cdots < i_k \leq m$.

Определение:

Диаграмма Юнга $\lambda = (\lambda_1, \ldots, \lambda_k)$ — это конечный невозрастающий набор чисел, задающий вид клетки Шуберта.

Решение: использование биективного сопоставления диаграмм Юнга и клеток Шуберта для параметризации грассманиана (Городенцев А.Л., 2011).
$\lambda_{k+1-j} = i_j - j$ при $j = (1, \ldots, k)$, где i_j — индекс последней единицы j-ой строки матрицы клетки Шуберта S_I.

Полученные результаты

Алгоритм перечисления точек грассманиана на основе использования диаграмм Юнга:

1. Генерируем всевозможные диаграммы Юнга.
2. Проводим сопоставление диаграмм матрицам коэффициентов клеток Шуберта.
3. Проводим перебор коэффициентов, оставшихся неопределенными.

Преимущества и отличия алгоритмов:

- Одна генерация диаграмм вместо k генераций кодов Грея.
- Прямое сопоставление матрицам вместо прямого перебора возможных вариантов в k вложенных циклах.

Результат: быстродействие алгоритма улучшено на 30%.

Дизайны и параметризация грассманиана

Определение:

Дизайн $D(v, b, r, k, \lambda)$ — размещение v элементов по b блокам размера k, где элемент встречается r раз, а пара — λ раз.

Идея: применение метода интегрирования дизайнов и рекуррентного порядка их построения для получения синдромов и соответствующей параметризации грассманиана.

Определение:

Отношение порядка $\prec$ согласовано с флагом F, если для $v \in V_i$, $\omega \in V_m \setminus V_i$ выполняется $v \prec \omega$.

Теорема о несогласованности с флагом:

Порядок на основе рекуррентных соотношений несогласован с флагом F на пространстве $V_m = (\mathbb{F}_q)^m$.

Вывод: метод неприменим для параметризации грассманиана.

Заключение

Результаты бакалаврской работы:

- Написана программа, реализующая точный критерий Фишера на языке Matlab.
- Реализован более быстрый алгоритм нахождения информативных признаков.
- Реализована программа нахождения наилучшего синдрома на языке R.
- Изучено применение рекуррентного порядка для параметризации грассманиана и доказано отсутствие его согласованности с флагом.
- Сделан вывод о факторах и взаимосвязях, влияющих на исход болезни.