Информационные методы идентификации значимых факторов

А.В. Булинский

(кафедра теории вероятностей мехмата
МГУ им. М.В.Ломоносова)

Конференция “Стохастика”
Математический институт им. В.А.Стеклова РАН,
15 октября 2024 года

Дорогой Альберт Николаевич!

С Днём рождения!

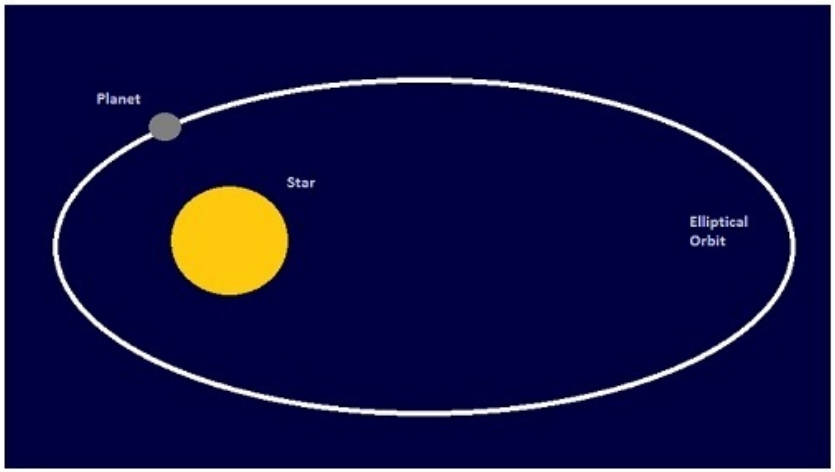
ПЛАН

1. Введение

Классическим примером анализа данных является открытие Й.Кеплером законов движения планет на основе наблюдений Т.Браге.

T.Brahe
(1546 - 1601)

J.Kepler
(1571 - 1630)



Первый закон Кеплера

Орбита планеты представляет собой эллипс, в одном из фокусов которого находится Солнце.

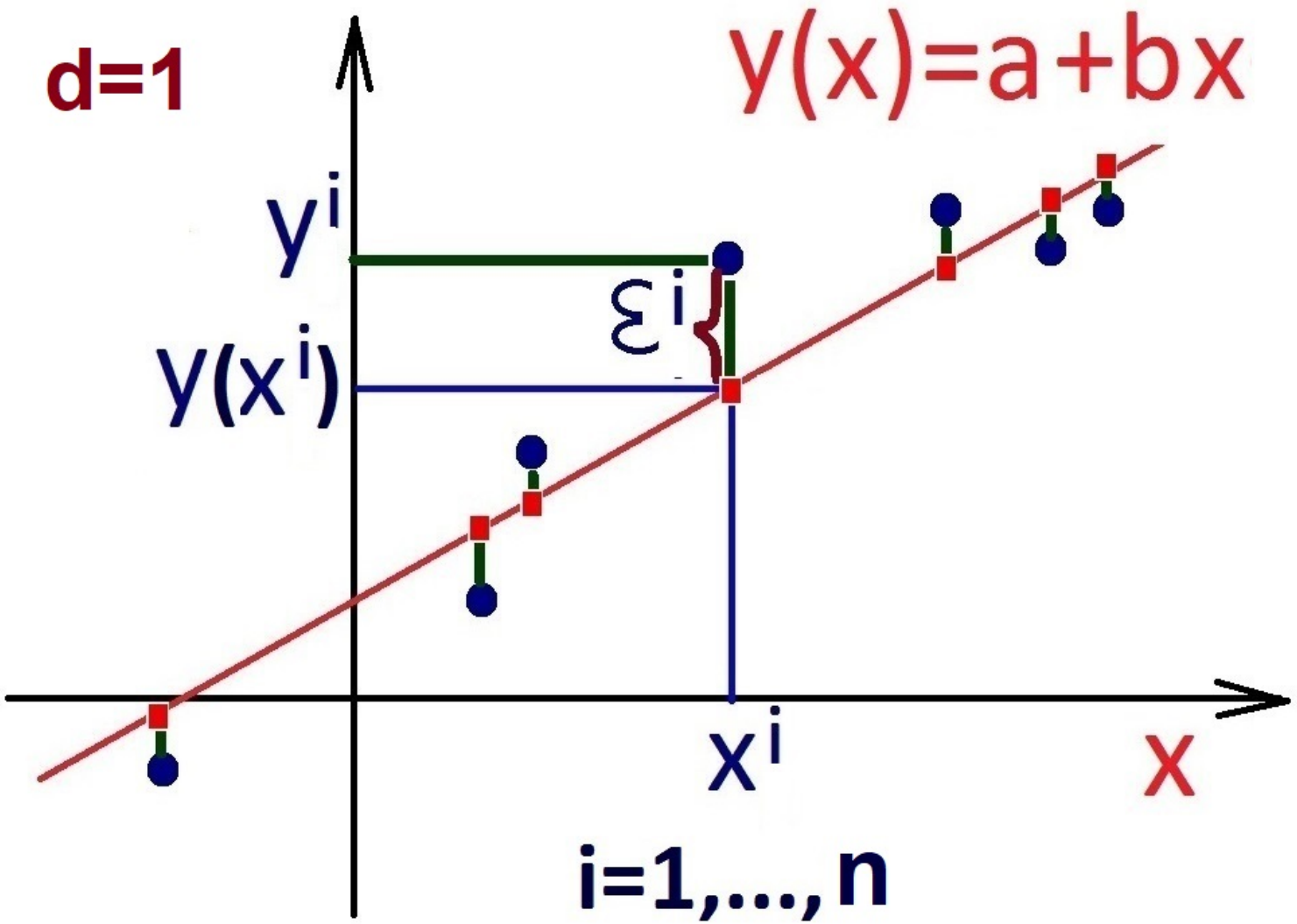
Во многих ситуациях интерес представляет переменная (отклик) Y и цель исследования – выявить зависимость между Y и набором факторов (признаков) $X = (X_1, \ldots, X_p)$. Предположим, что имеются наблюдения (X^i, Y^i) такие, что $X^i := (X_1^i, \ldots, X_p^i)$ – неслучайный вектор со значениями в $\mathbb{R}^p$, значения Y^i принадлежат $\mathbb{R}$, $i = 1, \ldots, n$. Наблюдения могут фиксироваться с ошибками. Иначе говоря, $Y^i = f(X^i) + \varepsilon^i$, где $f : \mathbb{R}^p \to \mathbb{R}$, $\varepsilon^1, \ldots, \varepsilon^n$ – н.о.р. величины с нулевым средним и дисперсией σ^2, $i = 1, \ldots, n$. Проблема заключается в нахождении функции f. Тогда истинное значение отклика, отвечающее X, будет $f(X)$.

Эта проблема восходит к А-М.Лежандру и К.Гауссу.

A-M. Legendre
(1752 - 1833)

K.F.Gauss
(1777 - 1855)

А-М.Лежанжр ввел в 1805 году "метод наименьших квадратов" (МНК) для решения указанной проблемы, когда f принадлежит определенному классу функций. В то время приложения этого метода относились к астрономии, точнее говоря, к изучению орбит планет.


d=1
y(x)=a+bx
y^i
ε^i
y(x^i)
x^i
x
i=1,...,n

В “Основах химии” Д.И.Менделеева (изд. 1906 г.) указано, что в 100 частях воды растворяется следующее число условных частей $NaNO_3$ при соответствующих температурах:

0°	4°	10°	15°	21°	29°	36°	51°	68°
66,7	71,0	76,3	80,6	85,7	92,2	99,4	113,6	125,1

Предполагается, что $y = ax + b$, где y – растворимость в условных частях на 100 частей воды, x – температура в градусах.

Д.И.Менделеев нашел в 1881 году, что $a = 67,5$; $b = 0,87$.

Работы А.А.Маркова (ученика П.Л.Чебышева) позволили в начале 20 века включить МНК в общую статистическую теорию оценивания.

Н.В.Маиевский опубликовал книгу “Traité de balistique exterieure”. Paris, Gauthier-Villars (1872), На с.XII написано: “При обработке опытов мы применили формулы Чебышева интерполяции по способу наименьших квадратов к определению проекции траектории снаряда на вертикальную плоскость стрельбы”.

P.Tchébychef. Formules d’interpolation par la méthode des moindres carrés. Mémoires couronnés et mémoires des savants étrangers. Acad.Roy.des Sciences,des Lettres et des Beau-Arts de Belg.,t.21, 1870, pp. 25-33.

Возникли РЕГРЕССИОННЫЕ МОДЕЛИ.

Линейная регрессия

$$Y = \sum_{i=1}^{p} a_i X_i + \varepsilon, \quad X_0 = 1.$$

Нелинейная регрессия

$$\mathrm{E}(Y|X) = g(X), \quad X = (X_1, \ldots, X_p),$$

где $g = g(x, \theta)$, $\theta \in \Theta$,
или $g \in \mathcal{G}$ (некоторый класс функций).
Существуют параметрические, непараметрические и полупараметрические регрессионные модели.

Напомним важную нелинейную модель
Пусть $Y : \Omega \to \{0, 1\}$. Например, 0 или 1 означают, что индивид здоров или болен.
Тогда $\mathbf{p}(x) := \mathsf{E}(Y|X = x) = \mathsf{P}(Y = 1|X = x)$, $x \in \mathbb{X} \subset \mathbb{R}^p$. Введем функцию

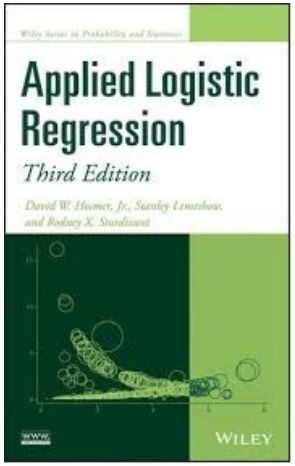
$$logit(t) := \log\left(\frac{t}{1-t}\right), \quad t \in (0, 1).$$

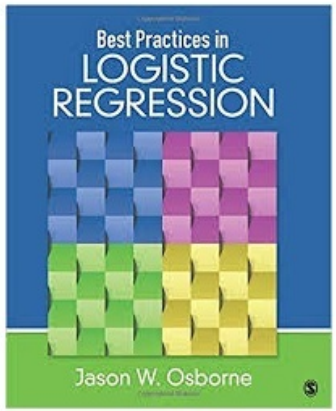
Логистическая регрессия определяется формулой

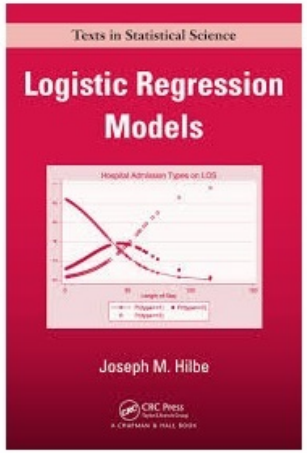
$$logit(\mathbf{p}(x)) = \sum_{i=0}^{p} a_i x_i.$$

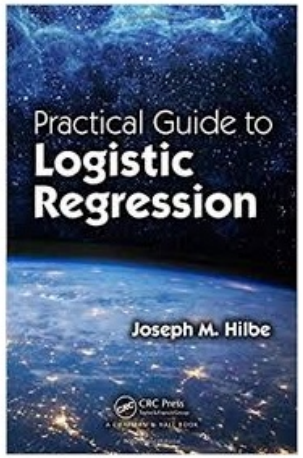
Иначе говоря, логистическая (сигмоидная) функция

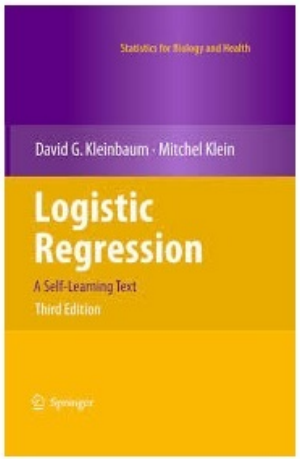
$$\mathbf{p}(x) = \frac{1}{1 + \exp\{-\sum_{i=0}^{p} a_i x_i\}}.$$

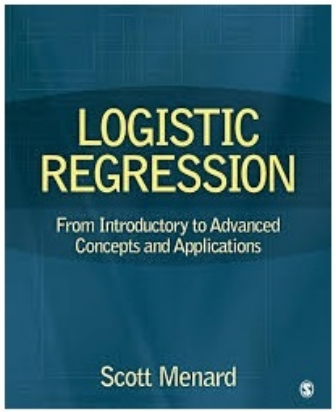


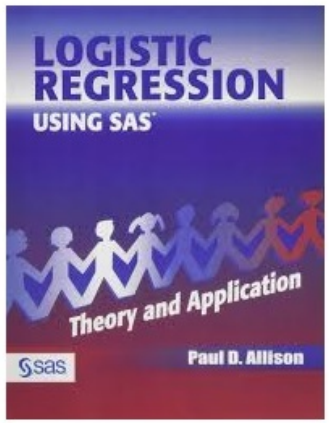


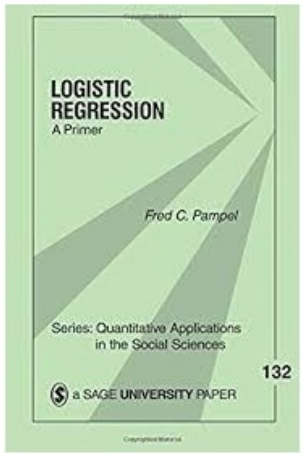










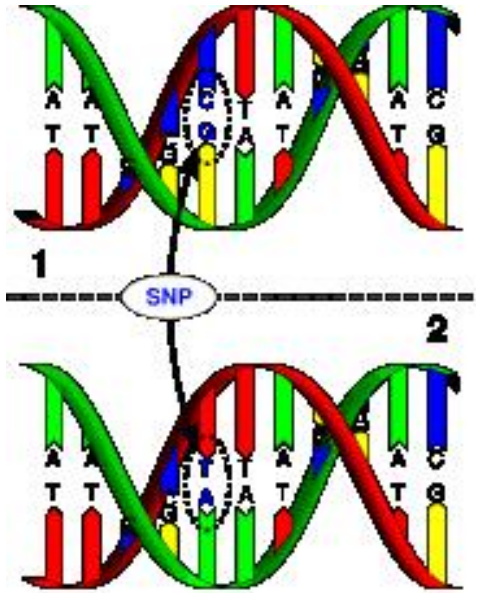


Книги о логистической регрессии

В наше время гигантские массивы наблюдений возникают при анализе микрочипов, классификации текстов, использовании высокочастотных финансовых данных, обнаружении интернет-мошенничеств и в других областях. Такие данные имеют огромные размерности (Big Data).

По прогнозам, к 2025 году в мире будет сгенерировано более 180 зеттабайт (или 180 триллионов гигабайт) данных. Данные по состоянию на 2015 год оцениваются в 10 зеттабайт. Наличие больших массивов данных стало причиной появления глубокого обучения, которое в дальнейшем привело к созданию искусственного интеллекта.

Каждая спираль генома человека содержит более трех миллиардов нуклеотидов. Только в 21-ом веке исследователям удалось прочитать такие гигантские последовательности “букв”, принадлежащих множеству $\{A, T, C, G\}$.



Теперь стало возможным определить положение генов и отдельных однонуклеотидных полиморфизмов (SNP) на упомянутой спирали.

"Поломки" (SNP) в структуре генома приводят к сложным заболеваниям (гипертензия, инфаркт миокарда, инсульт, диабет и другие). Проблема исключительной важности в генетике – идентифицировать наборы SNP, которые провоцируют сложные заболевания. В отличие от простых заболеваний (таких как серповидноклеточная анемия) влияние каждого SNP может быть мало, и только взаимодействие определенных повреждений структуры DNA может представлять опасность. Это называется эпистазом.

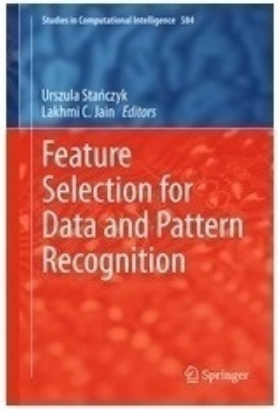
Простая XOR стохастическая модель, иллюстрирующая этот эффект, построена в статье A.Bulinski and A.Kozhevin (2018).

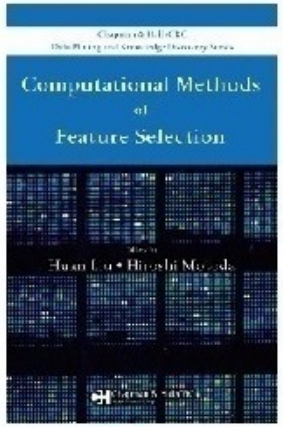
Выбор значимых факторов (Feature selection (FS)) – обширная область исследований, имеющая разнообразные приложения в медицине, биологии, финансах и других областях.

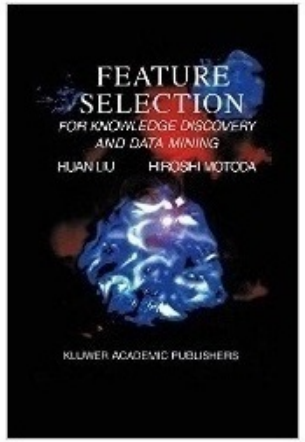
Среди переменных $X = (X_1, \ldots, X_p)$ ищется набор $X_S := (X_{i_1}, \ldots, X_{i_r})$, где $S = \{i_1, \ldots, i_r\} \subset \{1, \ldots, p\}$, такой, что X_S оказывает основное влияние (в определенном смысле) на поведение Y.
В отличие от классического МНК для построения модели используются не все факторы, а лишь часть из них. При этом важной является задача определения характера связи между Y и X_S.

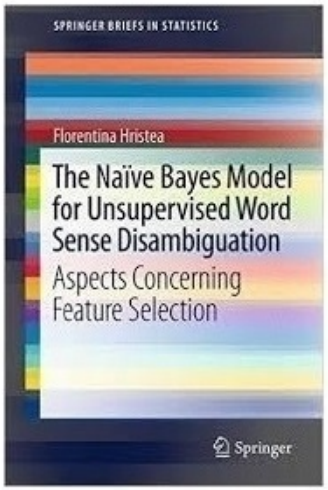
Продемонстрируем лишь некоторые книги, посвященные FS.

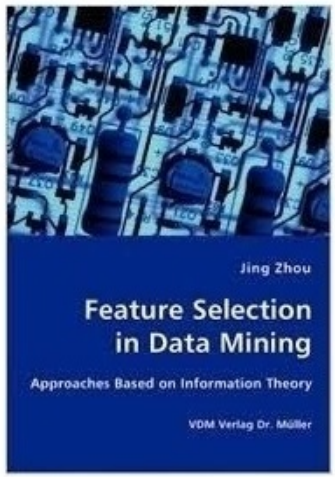

Verónica Bolón-Canedo
Noelia Sánchez-Maroño
Amparo Alonso-Betanzos
Feature Selection for High-Dimensional Data
Springer

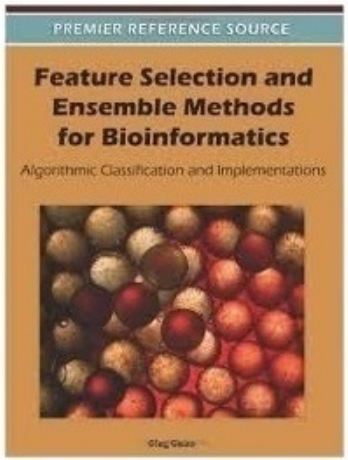

Urszula Stańczyk
Lakhmi C. Jain Editors
Feature Selection for Data and Pattern Recognition
Springer

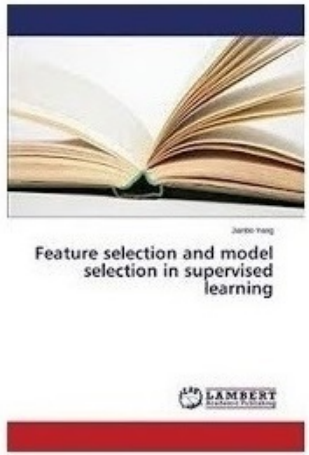

Computational Methods of Feature Selection


FEATURE SELECTION
FOR KNOWLEDGE DISCOVERY AND DATA MINING
HUAN LIU
HIROSHI MOTODA
KLUWER ACADEMIC PUBLISHERS

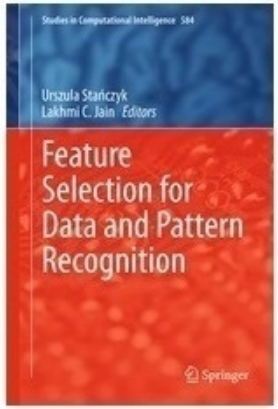

SPRINGER BRIEFS IN STATISTICS
Florentina Hristea
The Naïve Bayes Model for Unsupervised Word Sense Disambiguation
Aspects Concerning Feature Selection
Springer

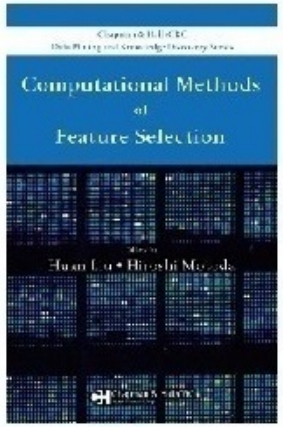

Jing Zhou
Feature Selection in Data Mining
Approaches Based on Information Theory
VDM Verlag Dr. Müller

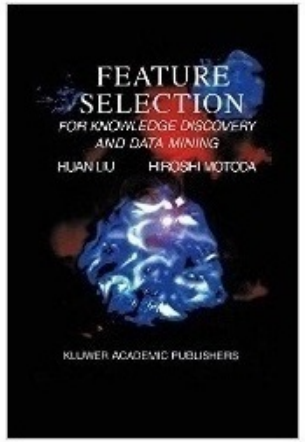

PREMIER REFERENCE SOURCE
Feature Selection and Ensemble Methods for Bioinformatics
Algorithmic Classification and Implementations

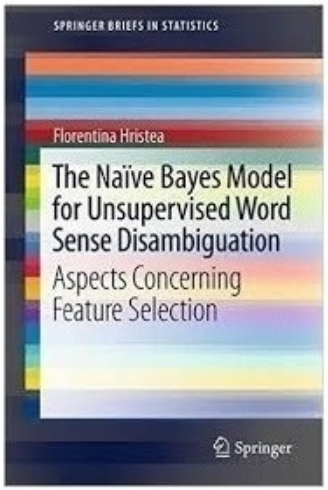

Feature selection and model selection in supervised learning
LAMBERT

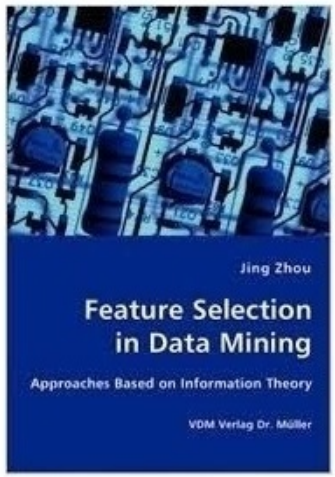

Verónica Bolón-Canedo
Noelia Sánchez-Maroño
Amparo Alonso-Betanzos
Feature Selection for High-Dimensional Data
Springer

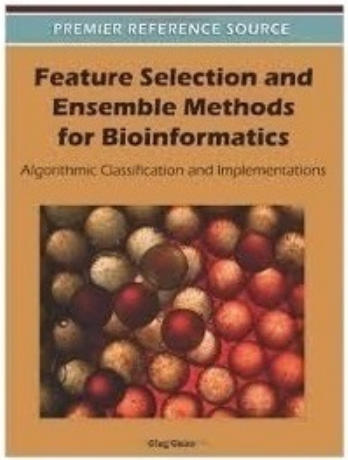

Urszula Stańczyk
Lakhmi C. Jain Editors
Feature Selection for Data and Pattern Recognition
Springer

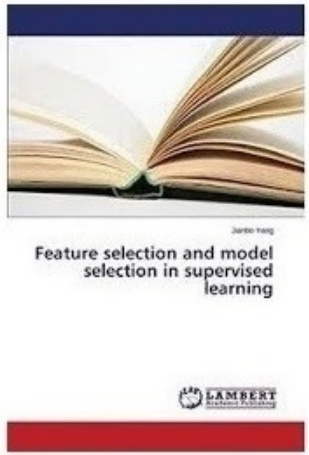

Computational Methods of Feature Selection

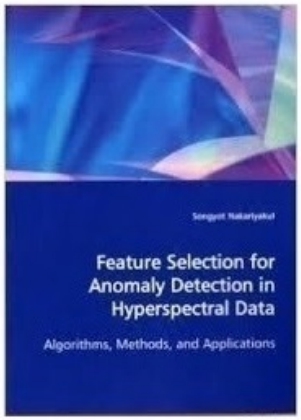

FEATURE SELECTION
FOR KNOWLEDGE DISCOVERY AND DATA MINING
HUAN LIU
HIROSHI MOTODA
KLUWER ACADEMIC PUBLISHERS

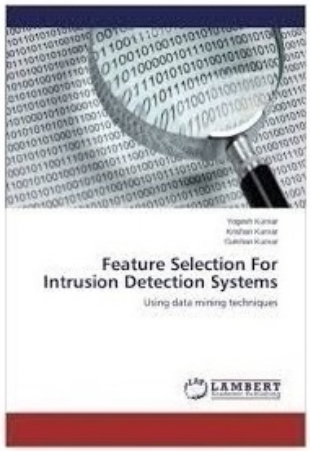

SPRINGER BRIEFS IN STATISTICS
Florentina Hristea
The Naïve Bayes Model for Unsupervised Word Sense Disambiguation
Aspects Concerning Feature Selection
Springer

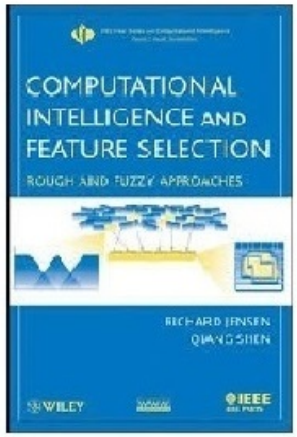

Jing Zhou
Feature Selection in Data Mining
Approaches Based on Information Theory
VDM Verlag Dr. Müller

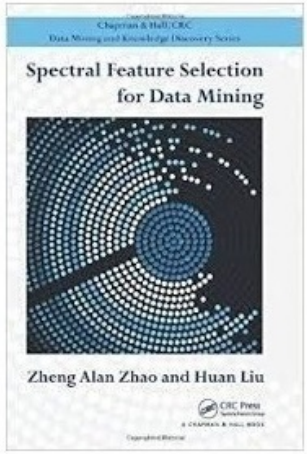

PREMIER REFERENCE SOURCE
Feature Selection and Ensemble Methods for Bioinformatics
Algorithmic Classification and Implementations

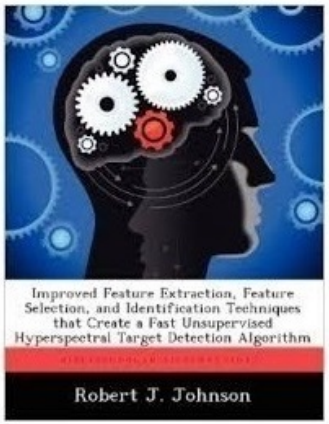

Feature selection and model selection in supervised learning
LAMBERT

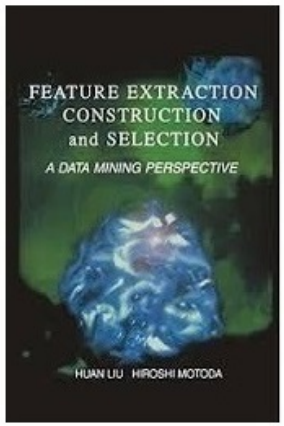

Feature Selection for Anomaly Detection in Hyperspectral Data
Algorithms, Methods, and Applications

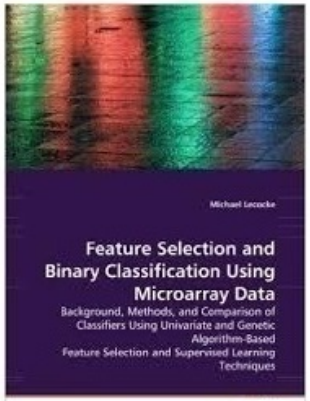

Feature Selection For Intrusion Detection Systems
Using data mining techniques
LAMBERT

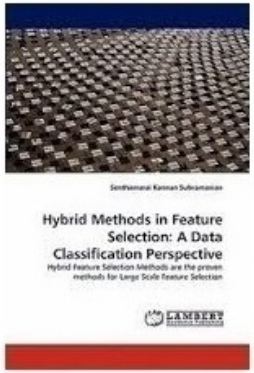

COMPUTATIONAL INTELLIGENCE AND FEATURE SELECTION
ROUGH AND FUZZY APPROACHES
RICHARD JENSEN
QIANG SHEN
WILEY
IEEE

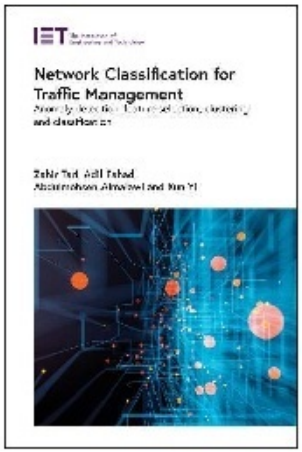

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
Spectral Feature Selection for Data Mining
Zheng Alan Zhao and Huan Liu
CRC Press


Improved Feature Extraction, Feature Selection, and Identification Techniques that Create a Fast Unsupervised Hyperspectral Target Detection Algorithm
Robert J. Johnson


FEATURE EXTRACTION CONSTRUCTION and SELECTION
A DATA MINING PERSPECTIVE
HUAN LIU HIROSHI MOTODA

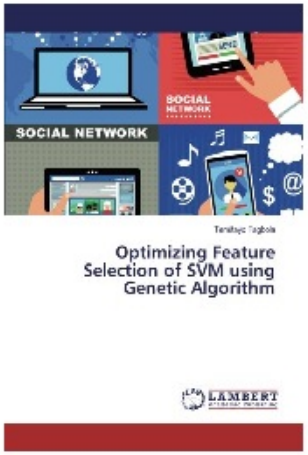

Michael Leyssche
Feature Selection and Binary Classification Using Microarray Data
Background, Methods, and Comparison of Classifiers Using Univariate and Genetic Algorithm-Based Feature Selection and Supervised Learning Techniques

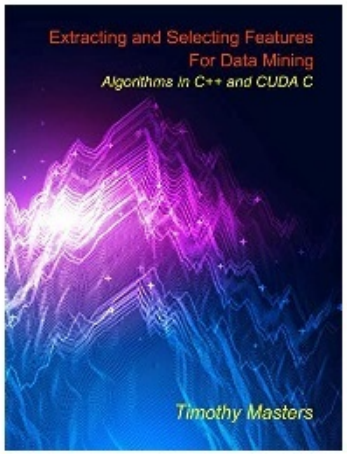

Hybrid Methods in Feature Selection: A Data Classification Perspective
LAMBERT

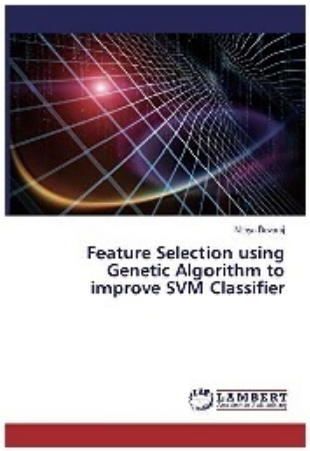

IET
Network Classification for Traffic Management

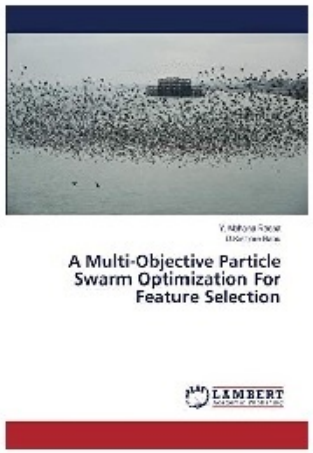

DATA SCIENCE SERIES
FEATURE ENGINEERING AND SELECTION
A Practical Approach for Predictive Models
MAX KUHN
KJELL JOHNSON

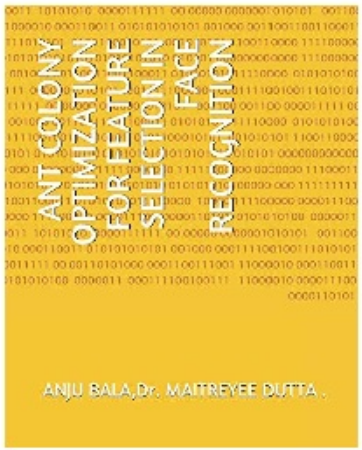

A quantum particle swarm optimization approach for feature selection
LAMBERT


SOCIAL NETWORK
SOCIAL NETWORK
Optimizing Feature Selection of SVM using Genetic Algorithm
LAMBERT


Extracting and Selecting Features For Data Mining
Algorithms in C++ and CUDA C
Timothy Masters


Feature Selection using Genetic Algorithm to improve SVM Classifier
LAMBERT


A Multi-Objective Particle Swarm Optimization For Feature Selection
LAMBERT


ANT COLONY OPTIMIZATION FOR FEATURE SELECTION IN FACE RECOGNITION
ANJU BALA,Dr. MAITREYEE DUTTA ,

Укажем также на ряд статей обзорного характера.

J.Li et al. (2018), L.Zhao et al. (2018), U.Stańczyk et al. (2018), K.Yu et al. (2020), P.Agrawal et al. (2021), E.O.Omuya, G.O.Okeyo, M.W.Kimwele (2021), N. Pudjihartono et al. (2022), P.Dhal, C.Azad (2022), N AlNuaimi (2022), D.Theng, K.K.Bhoyar (2023), Y.Lyu, Y.Feng, K.Sakurai (2023), H.H.Htun, M.Biehl, N.Petkov (2023), S.Sumanth (2023), C.Kuzudisli et al. (2023), R.Jiao et al. (2024).

FS может осуществляться с помощью ранжирования, взвешивания или выделения подмножества факторов. Ранжирование и взвешивание используются для выделения подмножества факторов с помощью определенного порога.

Наблюдения могут иметь или не иметь некоторые метки. Соответственно различают методы выбора факторов
с обучением (supervised methods),
полуобучающиеся (semi-supervised methods),
без обучения (unsupervised methods).

Ставятся различные задачи регрессии и классификации, а также кластеризации данных.

Выбор факторов осуществляется по трем причинам:

1) упрощение моделей, стремление сделать их интерпретируемыми для исследователей и пользователей;

2) сокращение времени на обучение (если оно производится);

3) чтобы избежать переподгонки модели.

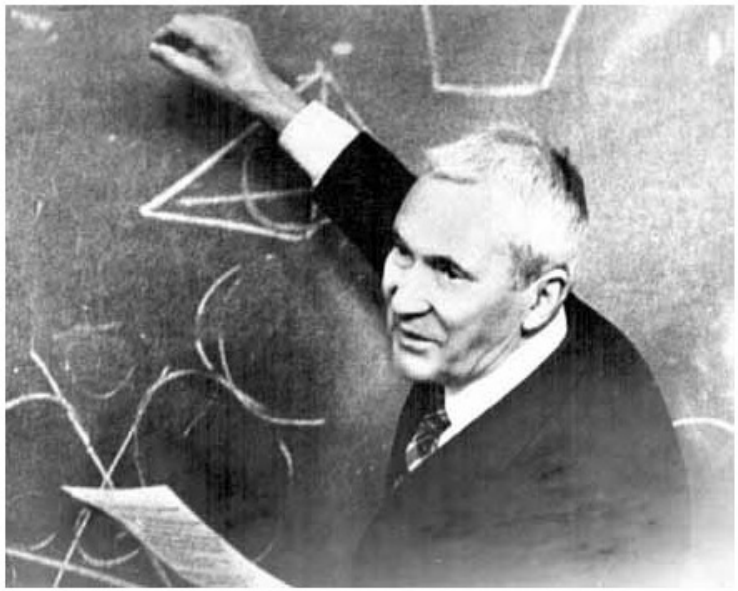
Обсуждение выбора факторов, обычно, концентрируется на двух аспектах:
стратегия выбора,
оценка ее качества.
Алгоритмы, применяемые при реализации намеченной стратегии,
делятся на следующие три категории:

фильтрующие (filter),
обертывающие (wrapper),
встраиваемые (embedded).

Встречаются также гибридные (hybrid) и усредняющие (ensemble) методы. Такие методы при анализе медицинских данных рассмотрены, например, в статье C-W.Chen et al. (2020).

Наряду с книгами имеется обширная журнальная литература в области отбора значимых факторов. Однако следует обратить внимание на то, что в подавляющем большинстве публикаций акцент делается не на доказательствах математических утверждений, а на введении новых алгоритмов и проведении компьютерных симуляций, иллюстрирующих эти алгоритмы.

Здесь уместно вспомнить слова великого ученого А.Н.Колмогорова:



А.Н.Колмогоров (1903 - 1987)

“Существует лишь тонкий слой между тривиальным и недоступным. В этом слое и делаются математические открытия”.

Таким образом, во второй половине 20-го века в математической статистике был сделан существенно новый шаг, который заключается в том, что вместо изучения всей выборки наблюдений стали интенсивно использоваться ее части. В этой связи отметим следующие направления исследований.

1. Выбор значимых факторов среди большого количества переменных.
2. Использование процедуры кросс-валидации.
3. Развитие бутстрэпа.

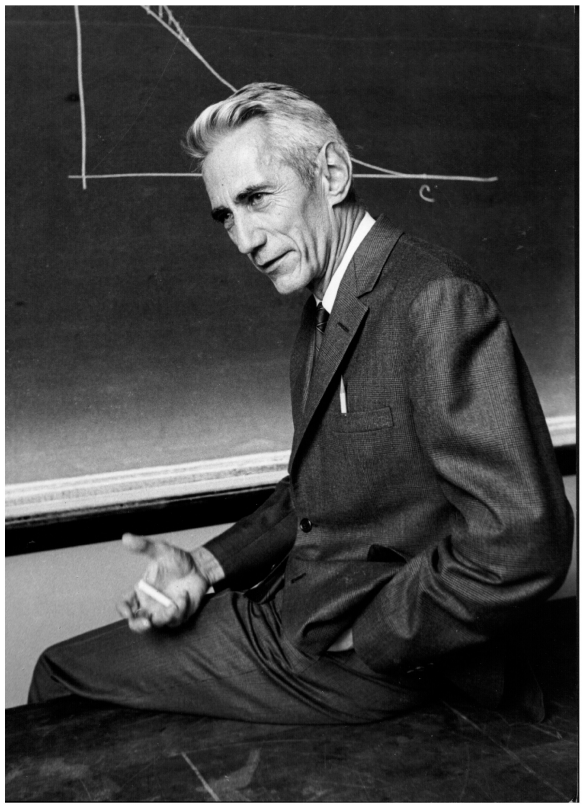
При этом компьютерное моделирование применяется и для иллюстрации теоретических результатов.

Вытаскивание из болота больших данных

2. Некоторые понятия теории информации

В отличие от использования коэффициента корреляции, направленного на выявление линейных зависимостей, теперь большое внимание уделяется применению основных понятий теории информации для идентификации нелинейных связей между случайными векторами. В этом отношении отметим энтропию, взаимную информацию, условную взаимную информацию и различные дивергенции, например, дивергенцию Kullback - Leibler.

Существуют разные (математические) определения энтропии. Напомним то, которое в 1948 году предложил



C.Shannon (1916 – 2001)

Если случайная величина X принимает значения в конечном или счетном множестве S, то энтропия Шеннона X (или распределения P_X величины X)

$$H(X) := -\sum_{x \in S} p_x \log p_x,$$

где $p_x := \mathrm{P}(X = x)$, $x \in S$, $0 \log 0 := 0$. Далее будем использовать натуральные логарифмы, поскольку постоянный множитель не играет роли.

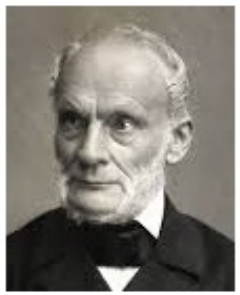
Пусть случайный вектор X принимает значения в $\mathbb{R}^d$ и имеет плотность f относительно меры Лебега μ, т.е.

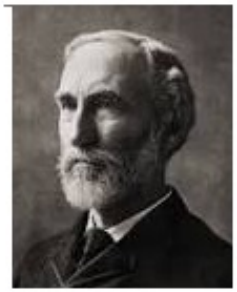
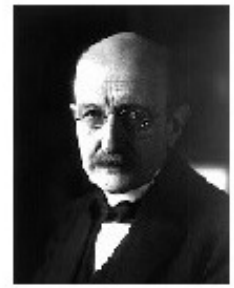
$$\mathrm{P}(X \in B) = \int_B f(x)\mu(dx),$$

где B – борелевское множество в $\mathbb{R}^d$.
(Дифференциальной) энтропией X или распределения с плотностью f называется

$$H(f) := -\int_{\mathbb{R}^d} f(x) \log f(x)\mu(dx).$$

Следует отметить выдающийся вклад, который внесли в развитие понятия энтропии R.Clausius (1822-1888), L.Boltzmann (1844-1906), J.Gibbs (1839-1903), M.Plank (1858-1947), C.Shannon (1916-2001), А.Н.Колмогоров (1903-1987), Я.Г.Синай и А.С.Холево.





Для вероятностных мер $\mathbb{P}$ и $\mathbb{Q}$ на пространстве $(S, \mathcal{B})$ дивергенция Кульбака - Лейблера

$$D_{KL}(\mathbb{P}||\mathbb{Q}) := \begin{cases} \int_S \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \text{если } \mathbb{P} \ll \mathbb{Q}, \\ \infty, & \text{иначе,} \end{cases}$$

где $\frac{d\mathbb{P}}{d\mathbb{Q}}$ – производная Радона - Никодима меры $\mathbb{P}$ по мере $\mathbb{Q}$.

Если $(S, \mathcal{B}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ и (абсолютно непрерывные) $\mathbb{P}$ и $\mathbb{Q}$ имеют плотности $p(x)$ и $q(x)$, $x \in \mathbb{R}^d$, относительно меры Лебега μ, то

$$D(\mathbb{P}||\mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mu(dx).$$

Формально полагаем $0/0 := 0$, $0 \cdot \log 0 := 0$.

Взаимная информация (MI) случайных элементов (векторов, величин) X и Y

$$I(X;Y) := D_{KL}(\mathrm{P}_{X,Y}||\mathrm{P}_X \otimes \mathrm{P}_Y),$$

где $\mathrm{P}_{X,Y}$ – совместное распределение (X, Y), а P_X и P_Y – соответственно распределения X и Y.
Имеем

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

Заметим, что $I(X;Y) \in [0, \infty]$, причем $I(X;Y) = 0$ тогда и только тогда, когда X и Y независимы.

Заметим, что существуют F-дивергенции между вероятностными мерами $\mathbb{P}$ и $\mathbb{Q}$, определенными на $(S, \mathcal{B})$. Следуя I.Csiszár (1963), T.Morimoto (1963) и S.M.Ali, S.D.Silvey (1966), положим для $\mathbb{P} \ll \mathbb{Q}$,

$$D_F(\mathbb{P}||\mathbb{Q}) := \int_S F\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q},$$

где $F : (0, \infty) \to \mathbb{R}$ – выпуклая функция,
$F(0) := \lim_{t \downarrow 0} F(t) \in (-\infty, \infty]$.
См, например, статью I.Sason (2022).

Энтропия, дивергенция Кульбака - Лейблера и взаимная информация широко используются в таких областях:

- статистические выводы (statistical inference)
- машинное обучение (machine learning)
- компьютерное зрение (computer vision)
- безопасность сетей (network security)
- выбор значимых факторов и классификация (feature selection and classification)
- физика
- биология
- финансы и др.

См., например, работы

P.Alonso-Ruiz and E.Spodarev (2019),
T.B.Berrett et al. (2019),
N.Carrara and J.Ernst (2019),
A.Charzyńska and A.Gambin (2016),
C-A.Deledalle (2017),
C.Granero-Belinchón et al.(2018),
J.Li et al. (2017),
T.Ma et al.(2016),
K.R.Moon et al. (2019),
P.Moulin and V.V.Veeravalli (2019),
L.Pardo (2019).
J.Suzuki (2021).

Следуя работам W.J.McGill, (1954), T.S.Han (1980) и R.M.Fano (1961), определим информацию взаимодействия величин $X_1, \ldots, X_m$, обобщающую взаимную информацию для двух величин,

$$II(X_1, \ldots, X_m) = I(\{X_1, \ldots, X_m\})$$
$$:= -\sum_T (-1)^{|T|} H(X_T),$$

где сумма берется по всем подмножествам T множества $\{1, \ldots, m\}$ ($H_\varnothing := 0$). Введем условную информацию взаимодействия

$$II(X_1, \ldots, X_m | X_{m+1}) := -\sum_T (-1)^{|T|} H(X_T | X_{m+1}).$$

Напомним, что для дискретных величин U и V условная энтропия задается формулой

$$H(U|V) = -\sum_{j=1}^{m}\sum_{i=1}^{q} p(u_i, v_j)\log p(u_i|v_j),$$

где $p(u_i, v_j) := \mathrm{P}(U = u_i, V = v_j)$ и $p(u_i|v_j) := \mathrm{P}(U = u_i|V = v_j)$, $i = 1, \ldots, q$, $j = 1, \ldots, m$. Имеем

$$H(U|V) = H(U, V) - H(V),$$

$$I(U; V) = H(U) - H(U|V),$$

$$I(U; V) = I(V; U) = H(V) - H(V|U).$$

Если ставится задача найти набор из r факторов ($r < p$), который является "наиболее информативным" для описания Y, то выбирается X_{S_0} такое, что

$$I(X_{S_0}; Y) = \max_{S \subset \{1,\ldots,p\},\ |S|=r} I(X_S; Y).$$

Не предполагается, что S_0 определено однозначно. Нахождение набора X_{S_0} было названо в статье H.Peng, F.Long, C.Ding (2005) "схемой максимальной зависимости" ("Max-Dependency scheme").

Указанный выбор S_0 является полезным и в силу формулы (1) из статьи G.Brown (2009) для байесовской ошибки бинарной классификации:

$$\mathrm{P}(g(V) \neq U) \leq \frac{1}{2} H(U|V),$$

где g – произвольная измеримая функция, отображающая область значений V в область значений U. Эта формула приписывается ряду авторов. По-видимому, впервые она установлена в работе В.А.Ковалевского (1968).

Для случайной величины, принимающей конечное число значений, энтропия неотрицательна, поэтому и $H(U|V) \geq 0$, а так как $I(U,V) \geq 0$, то для рассматриваемой величины U значение $H(U|V)$ будет тем меньше, чем ближе $I(U;V)$ окажется к $H(U)$. В частности, $H(U|U) = 0$.
При аппроксимации U с помощью V естественно возникает задача максимизации $I(U;V)$.
Следовательно, для бинарной величины Y выбор множества S_0, гарантирует минимизацию верхней оценки ошибочной классификации Y с помощью некоторой функции $g(X_S)$. Заметим также, что неравенство Фано ((6.16) в книге Р.Фано (1965)) и его обобщения дают нижние оценки для $\mathrm{P}(g(V) \neq U)$.

В многочисленных публикациях авторы справедливо указывают, что максимизация функции, заданной на всех подмножествах множества $\{1, \ldots, p\}$, имеющих мощность r, практически нереализуема при весьма больших p и умеренно больших r.

Перебор всех подмножеств даже из трех элементов, описывающих места “поломок” генома, не реализуем с помощью современных компьютеров за обозримое время. Считается, что целесообразно рассматривать только некоторые определенные участки молекулы ДНК, связанные с данным сложным заболеванием. Тогда задача нахождения S_0 становится вычислительно выполнимой.

Трудности выбора значимых факторов обусловлены также тем, что совместное распределение (X, Y), вообще говоря, неизвестно. Естественно использовать аналог экстремальной задачи нахождения S_0, где вместо $I(X_S; Y)$ применяется ее статистическая оценка $\widehat{I}_n(X_S; Y)$, построенная по наблюдениям $(X_1^{(j)}, \ldots, X_p^{(j)}, Y^{(j)})$, $j = 1, \ldots, n$. Обычно предполагается, что $(X_1^{(j)}, \ldots, X_p^{(j)}, Y^{(j)})_{j \in \mathbb{N}}$ – последовательность независимых одинаково распределенных случайных векторов, имеющих такое же распределение, как (X, Y).

Следовательно, предварительно требуется найти статистические оценки интересующих характеристик, а затем решать упомянутую экстремальную задачу, опираясь на эти оценки. Статистическое оценивание информационных характеристик – нетривиальная задача. В этой области используются разнообразные методы, например, статистические оценки, строящиеся на основе техники ближайших соседей (см. монографию G.Biau, L.Devroye (2015)), аппарат условных распределений, анализ равномерной интегрируемости семейства функций и др.). В этой связи укажем на наши следующие работы, в которых можно найти обширную библиографию.

A.Bulinski and D.Dimitrov. Statistical estimation of the Shannon entropy. Acta Mathematica Sinica, English Series, 2019, 35, p. 17-46.

A.Bulinski and A.Kozhevin. Statistical estimation of conditional Shannon entropy. ESAIM - Probability and Statistics, 2019, 23, p. 350-386.

A.Bulinski and A.Kozhevin. Statistical Estimation of Mutual Information. Methodlology and Computing in Applied Probability, 2021, 23, p. 123-142.

A.Bulinski and D.Dimitrov. Statistical estimation of the Kullback-Leibler divergence. Mathematics, 2021, v.9, 544, p. 1-36.

Таким образом, неудивительно, что появилось много субоптимальных процедур нахождения значимых факторов, влияющих на переменную отклика. Достаточно указать на прямую и обратную процедуры отбора признаков. Анализ прямых методов отбора, использующих взаимную информацию, проведен в статье F.Macedo (2019), см. также обзор J.Mielniczuk (2022).

Мы покажем, что широко применяемые субоптимальные алгоритмы выбора значимых факторов, основанные на понятиях теории информации, могут лишь с малой вероятностью идентифицировать оптимальный набор S_0. Это отражает эффект эпистаза (взаимодействия признаков).

Таким образом, в общей последовательной процедуре за r шагов выбираются факторы с попарно различными номерами $i_1, \ldots, i_r$, где $S = \{i_1, \ldots, i_r\} \subset \{1, \ldots, p\}$. Также имеется подход последовательного исключения из всей совокупности p факторов набора из $p - r$ факторов, чтобы оставить r (в определенном смысле) значимых. Мы сконцентрируемся на процедурах последовательного включения признаков ("forward selection"). Вначале не будем рассматривать статистические оценки используемых функционалов. Позже вернемся к этому вопросу.

Как правило, на первом шаге выбирается признак с номером i_1 таким, что

$$I(X_{i_1}; Y) = \max_{i=1,\ldots,p} I(X_i; Y).$$

Всюду далее множество $S_{t-1} = \{i_1, \ldots, i_{t-1}\} \subset T$ обозначает набор индексов, построенный за первые $(t-1)$ шагов, где $t \geq 2$ и $T = \{1, \ldots, p\}$. Добавление к S_{t-1} элемента i_t производится на основании максимизации некоторого функционала. Простейший способ заключается в том, чтобы выбирать i_t, для которого

$$I(X_{i_t}; Y) = \max_{i \in T \setminus S_{t-1}} I(X_i; Y),$$

где $t = 2, \ldots, r$, $r < p$. Если таких i несколько, то из них берется любая с равной вероятностью.

В статье G.Brown (2009) предложен общий подход к построению субоптимальных алгоритмов последовательного отбора факторов, основанный на установленной в указанной работе теореме 1. А именно, эта теорема утверждает (в наших обозначениях), что для любого множества $S \subset \{1, \ldots, p\}$ справедлива формула

$$I(X_S; Y) = \sum_{W \subset S} I(\{X_W, Y\}),$$

где в сумме фигурирует информация взаимодействия. В качестве аппроксимации этой суммы можно ограничиться слагаемыми, для которых $|W| \leq w$ при некотором $w \in \mathbb{N}$.

Положим

$$I^{(w)}(X_S; Y) := \sum_{W \subset S,\, |W| \le w} I(\{X_W, Y\}).$$

Для заданного r $(1 \le r < p)$ рассмотрим ту же задачу нахождения множества S_0, заменив I на $I^{(w)}$. Тогда существует хотя бы одно множество $S_0 = S_0(r, w)$, являющееся решением упомянутой задачи. Например, если $w = 1$, то для $S = \{k_1, \ldots, k_r\}$ имеем

$$I^{(1)}(X_S; Y) = \sum_{i \in S} I(\{X_i, Y\}) = \sum_{t=1}^{r} I(X_{k_t}; Y).$$

Поэтому для получения $S_0(r,1)$ ранжируем числа $I^{(1)}(X_1;Y),\ldots,I^{(1)}(X_p;Y)$. Получаем набор вида $a_1 \le a_2 \le \ldots \le a_p$ и берем среди них $a_{p-r+1},\ldots,a_p$, что определяет соответствующие индексы, входящие в множество $S_0(r,1)$ (когда среди $a_1,\ldots,a_p$ имеются совпадающие числа, то выбор $S_0(r,1)$ может оказаться неоднозначным). Заметим, что этот набор индексов заведомо совпадает с тем, который находился с помощью введенного выше последовательного алгоритма, если процедура выбора $S_0(r,1)$ с помощью $I^{(1)}$ однозначна.

Вероятно, впервые в работе R.Battiti (1994) был предложен подход, который, в отличие от описанного выше алгоритма выбора r признаков на основе функционала $I^{(1)}$, также учитывал, сколь связанными друг с другом оказывались выбираемые признаки. При этом во многих работах авторы стремились найти некоторый баланс между информативностью выделяемого набора признаков и содержащейся в нем "избыточностью информации" (когда практически та же информация имеется в наборе мало связанных друг с другом признаков).

Для этого R.Battiti был введен функционал, с помощью которого на шаге t выбирался признак X_{i_t} (т.е. индекс i_t), максимизировавший величину

$$J_{MIFS}(X_i) := I(X_i; Y) - \beta \sum_{k \in S_{t-1}} I(X_i; X_k)$$

по $i \in \{1, \ldots, p\} \setminus S_{t-1}$. Здесь β – некоторый параметр, определяющий вес штрафной функции, учитывающей парные взаимодействия фактора, выбираемого на шаге t, с уже выбранными ранее (MIFS обозначает Mutual Information Feature Selection).

Функционал вида J_{MIFS} был независимо использован в статье H.Peng et al. (2005) для последовательного отбора признаков при $\beta = \beta(t) = \frac{1}{t-1}$, где $t \geq 2$. Соответствующий алгоритм получил название mRMR (minimum Redundancy Maximum Relevance), т.е. “минимум избыточности, максимум значимости”. Отметим также, что различные варианты функционалов, учитывающие упомянутый выше баланс, были предложены в ряде работ, ссылки на которые приведены в статье G.Brown (2009).

Для $w = 2$ имеем

$$I^{(2)}(X_S; Y) = \sum_{k \in S} I(\{X_k, Y\}) + \sum_{k,v \in S,\ k<v} I(\{X_k, X_v, Y\})$$

$$= \sum_{k \in S} I(X_k; Y) + \sum_{k,v \in S,\ k<v} (I(X_k; X_v) - I(X_k; X_v | Y)),$$

где $I(X_k; X_v | Y) := I(\{X_k, X_v\} | Y)$.

В работе G.Brown выбор каждого признака на шаге t осуществлялся с помощью максимизации функционала

$$J(X_i) := I(X_i; Y) - \beta \sum_{k \in S_{t-1}} I(X_i; X_k) + \gamma \sum_{k \in S_{t-1}} I(X_i; X_k | Y)$$

по $i \in \{1, \ldots, p\} \setminus S_{t-1}$, где $t \geq 2$, а β и γ – некоторые числовые параметры. В частности, при $\gamma = 0$ получаем, что $J = J_{MIFS}$.

В статье H.Peng et al. (2005), наряду с Max-Dependency алгоритмом построения множества значимых факторов, описан и последовательный вариант Max-Dependency алгоритма, когда первый шаг означал что $I(X_{i_1}; Y) = \max_{i=1,\ldots,p} I(X_i; Y)$, а на шаге t выбирался признак X_{i_t}, который максимизировал функционал

$$J(X_i) := I(X_{S_{t-1}}, X_i; Y)$$

по всем $i \in \{1, \ldots p\} \setminus S_{t-1}$, $t \geq 2$ ("Max-Dependency first order incremental search", когда на каждом шаге один признак добавляется к уже построенным).

Кроме того, в той же статье предложен алгоритм наибольшей значимости (“Max-Relevance”), который означал максимизацию $\frac{1}{|S|}\sum_{i\in S} I(X_i; Y)$ по всем $S \subset \{1, \ldots, p\}$, для которых $|S| = r$. Очевидно, эта процедура равносильна описанному выше применению $I^{(1)}$. Однако основное внимание авторы уделили алгоритму mRMR, в котором последовательная процедура отбора факторов (или номеров факторов) осуществляется с помощью функционала J_{MIFS} при $\beta_t = \frac{1}{t-1}$, $t \geq 2$. Если максимум достигается в нескольких точках, то из них с равной вероятностью выбирается любая.

Главный результат работы Peng et al. (2005) (теорема на с. 3) утверждает, что mRMR алгоритм и (последовательный) Max-Dependency алгоритм приводят к одному и тому же набору факторов, когда все используемые значения величин взаимной информации известны. Имеется более 10 000 ссылок на указанную статью, положившую начало широкому применению метода mRMR. Однако в 2020 была опубликована статья P.Bugata, P.Drotar, в которой авторы продемонстрировали, что этот знаменитый результат некорректен. Подчеркнем, что важная заслуга H.Peng и соавторов заключается в развитии теоретического направления анализа качества субоптимальных процедур.

В статье P.Bugata, P.Drotar (2020) использовался случайный вектор (X_1, X_2, X_3, Y) такой, что mRMR алгоритм для $r = 2$ приводил к набору (X_1, X_3) в то время как алгоритм последовательного поиска с максимальной зависимостью (Max-Dependency) давал (X_1, X_2). В построенном авторами примере $I(X_1, X_2; Y) > I(X_1, X_3; Y)$. Это противоречит утверждению теоремы из работы H.Peng et al. (2005). Этот пример в несколько ином изложении, а также его обобщение, подробно рассмотрен в статье А.В.Булинского (2023).

Итак, классическая версия mRMR и Max-Dependency алгоритмов могут приводить к различным наборам из r "значимых факторов". Заметим, что только случай $r = 2$ рассматривался в статье P.Bugata, P.Drotar (2020), причем авторы оперировали значениями взаимной информации, которые считались известными. Статистические оценки этих характеристик не использовались. Теперь мы обратимся к более общей ситуации и для каждого $r \geq 2$ покажем, что множества признаков, построенные с помощью последовательного алгоритма Max-Dependency (или mRMR) и алгоритма, основанного на наиболее информативном поиске, не обязаны совпадать.

Лемма (А.В.Булинский (2023))

Пусть независимые случайные величины $X_1, \ldots, X_p$ *принимают значения* 1 *и* -1 *с вероятностью* $\frac{1}{2}$, *где* $p \in \mathbb{N}$. *Возьмем любое непустое множество* $T \subset \{1, \ldots, p\}$ *и введем*

$$X_{p+1} := \prod_{k \in T} X_k.$$

Тогда для любого множества $J \subset \{1, \ldots, p+1\}$ *случайные величины* $X_j, j \in J$, *являются зависимыми в том и только том случае, когда* $T \cup \{p+1\} \subset J$ (“$\subset$” *обозначает* “$\subseteq$”).
В частности, для $I \subset \{1, \ldots, p+1\}$ *набор* $\{X_j, j \in I\}$ *состоит из независимых одинаково распределенных величин, когда* $|I| \leq |T|$.

Теорема (А.В.Булинский (2023))

Пусть $r \in \mathbb{N}$ *и* $p \in \mathbb{N}$ *таковы, что* $2 \leq r < p$. *Тогда существует набор факторов* $X_1, \ldots, X_p$ *и отклик* Y *такие, что Max-Dependency алгоритм последовательного отбора факторов, задаваемый функционалом* $I(X_{S_{t-1}}, X_i; Y)$, *с вероятностью* $r/\binom{p}{r-1}$ *идентифицирует набор наиболее значимых признаков* X_{S_0}, *где* $|S_0| = r$. *Для тех же случайных величин* $X_1, \ldots, X_p, Y$ *алгоритм mRMR правильно идентифицирует множество* S_0 *с вероятностью* $1/\binom{p}{r}$.

Необходимо подчеркнуть еще раз, что в практических задачах закон распределения (X, Y) неизвестен, и мы используем статистические оценки $\widehat{I}_n(X_S; Y)$ взаимной информации $I(X_S; Y)$, где $S \subset \{1, \ldots, p\}$. Такие оценки обычно строятся при помощи независимых наблюдений $(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots$, имеющих такое же распределение, как (X, Y), см., например, гл. 5 в книге Z.Zhang (2017), статью A.Bulinski, A.Kozhevin (2021) и там же ссылки. Возникает задача нахождения

$$\widehat{S}_0^{(n)} \in \arg \max_{S \subset \{1,\ldots,p\}, |S|=r} \widehat{I}_n(X_S; Y),$$

где $\widehat{S}_0^{(n)} = \widehat{S}_0^{(n)}(r)$.

Теперь при использовании в нашем примере (лемма 1) вариантов последовательного алгоритма Max-Dependency или mRMR, основанных на статистических оценках применяемых функционалов, уже на первом шаге мы сталкиваемся не с набором нулей (так как $I(X_i; Y) = 0$ для всех $i = 1, \ldots, p$), а с набором одинаково распределенных случайных величин

$$Z_{n,j} := \widehat{I}_n(X_j; Y), \quad j = 1, \ldots, p, \quad n \in \mathbb{N}.$$

Для иллюстрации сказанного обратимся к простейшему варианту оценки взаимной информации. Иначе говоря, рассмотрим статистические оценки “непосредственной подстановки” (“plug-in”), когда оценка значения некоторой функции $h(x_1, \ldots, x_q)$ имеет вид $h(\widehat{x}_1, \ldots, \widehat{x}_q)$, где $\widehat{x}_i$ – определенная статистическая оценка величины x_i, $i = 1, \ldots, q$.

Пусть независимые случайные величины $X_1, \ldots, X_p$ принимают значения -1 и 1 с вероятностью $1/2$, где $p \in \mathbb{N}$. Возьмем любое непустое подмножество $T \subset \{1, \ldots, p\}$ и введем отклик $Y = \prod_{k \in T} X_k$, где $|T| = r \geq 2$.
С помощью леммы 1 нетрудно видеть, что для так заданных случайных величин $X_1, \ldots, X_p$ и Y множество $S_0 = S_0(r)$ совпадает с T. Рассмотрим

$$X_{n,j} = (X_j^{(1)}, \ldots, X_j^{(n)}), \ W = (W_1, \ldots, W_n),$$

$$W_t = \prod_{k \in T} X_k^{(t)}, \ n \in \mathbb{N}, \ j = 1, \ldots, p, \ t = 1, \ldots, n.$$

Разумеется, $W = W(n)$. Введем оценку подстановки ("plug-in") взаимной информации величин X_m и Y, построенную по векторам $(X_{n,m}, W)$:

$$Z_{n,m} := \widehat{I}_n(X_m, Y), \quad n \in \mathbb{N}, \ m = 1, \ldots, p,$$

где для $u_i, v_j \in \{-1, 1\}$, $i, j = 1, 2$, полагаем

$$\widehat{I}_n(X_m, Y) = \sum_{i=1}^{2} \sum_{j=1}^{2} \widehat{p}_n(u_i, v_j) \log \frac{\widehat{p}_n(u_i, v_j)}{\widehat{p}_n(u_i)\widehat{q}_n(v_j)},$$

$$\widehat{p}_n(u_i, v_j) := \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\{X_m^{(t)} = u_i, W_t = v_j\},$$

$$\widehat{p}_n(u_i) = \sum_{j=1}^{2} \widehat{p}_n(u_i, v_j), \ \widehat{q}_n(v_j) = \sum_{i=1}^{2} \widehat{p}_n(u_i, v_j).$$

Как обычно, формально полагаем $0/0 := 1$ и $0 \log 0 := 0$. Для упрощения обозначений не пишем, что функции $\widehat{p}_n(u_i, v_j)$, а следовательно, $\widehat{p}_n(u_i)$ и $\widehat{q}_n(v_j)$ зависят от m. Таким образом, $Z_{n,m} = \varphi(X_{n,m}; W)$, где φ – борелевская функция, зависящая от n.

На первом шаге последовательного отбора факторов берется точка j_1, которая обеспечивает максимум величин $\widehat{I}_n(X_j; Y)$ по всем $j \in \{1, \ldots, p\}$. Если таких точек несколько, то с равной вероятностью выбирается любая из них. Обозначим $P(n)$ вероятность того, что j_1 принадлежит множеству T, когда для ее нахождения использовались векторы $(X^{(j)}, W_j)$, $j = 1, \ldots, n$.

Теорема (А.В.Булинский (2024))

В рамках описанной выше модели при $r \geq 2$ *и при* $p \geq 2r - 1$ *справедливо соотношение*

$$P(n) \leq \frac{1}{k+1} + \frac{1}{p-r+1} + O(\frac{1}{\sqrt{n}}), \ n \to \infty,$$

где $r = |T|$ *и* $k = [(p-r)/(r-1)]$, *а* $[\cdot]$ *обозначает целую часть числа.*

Доказательство использует следующий результат, представляющий самостоятельный интерес.

Лемма (А.В.Булинский (2024))

Пусть функция $g : [0,1] \to \mathbb{R}$ *не убывает, а дискретная случайная величина* Z *принимает значение* x_i *с вероятностью* p_i *, где* $i = 1, \ldots, v$. *Тогда*

$$\mathrm{E}g(F(Z)) \leq \int_0^1 g(\min\{p + a, 1\}dp,$$

где F *– функция распределения* Z *и* $a = \max_{i=1,\ldots,v} p_i$.

Замечание

Утверждение теоремы 1 сохранится, если выбор признака на шаге $t = 1, \ldots, r$ осуществляется с помощью максимизации функционала

$$G_t(X_i) := G(X_i, Y) - \beta_t \sum_{k \in S_{t-1}} G(X_i, X_k),$$

где $i \in \{1, \ldots, p\} \setminus S_{t-1}$ ($S_0 := \varnothing$, сумма по пустому множеству считается равной нулю), $\beta_1, \ldots, \beta_r$ – произвольные числовые параметры, G такой коэффициент зависимости, что $G(U, V) = 0$, когда случайные векторы U и V независимы. Более того, вывод сохранится, если вместо $G(X_i, Y)$ записать $G((X_{S_{t-1}}, X_i), Y)$.

Теорема (А.В.Булинский (2023))

Пусть $r \in \mathbb{N}$ *и* $p \in \mathbb{N}$ *таковы, что* $2 \leq r < p$, *а алгоритм последовательного отбора факторов основан на максимизации на шаге* t *функционала*

$$J_t(X_i) := I(X_i; Y) - \beta_t \sum_{k \in S_{t-1}} I(X_i; X_k)$$

$$+\gamma_t \sum_{k \in S_{t-1}} I(X_i; X_k | Y)$$

по $i \in \{1, \ldots, p\} \setminus S_{t-1}$, β_t *и* γ_t *являются некоторыми числовыми параметрами,* $t = 1, \ldots, r$. *Тогда существует набор факторов* $X_1, \ldots, X_p$ *и отклик* Y *такие, что при* $2 < r < p$ *указанный алгоритм с вероятностью* $1/\binom{p}{r}$ *идентифицирует набор значимых признаков* S_0. *Если* $2 = r < p$, *то вероятность идентификации* T *равна* $1/p$.

Теорема (А.В.Булинский (2023))

Пусть натуральные r, p и w таковы, что $2 \leq r < p$ и $1 \leq w < p$. Тогда утверждение предыдущей теоремы имеет место, если на шаге $t \in \{1, \ldots, r\}$ вместо J_t используется функционал $J_t^{(w)}$.

Таким образом, представляется целесообразным сконцентрироваться на моделях с определенными связями между переменной отклика Y и факторами $X_1, \ldots, X_p$. В работе A.A.Kozhevin (2021) исследовалась смешанная модель и для нее установлена процедура состоятельного оценивания множества S_0 при заданном r.

4. Дельта-метод

Дельта метод имеет длинную историю, восходящую к исследованиям 18-го века. В его развитие внесли вклад такие известные ученые как C.F.Gauss, C.E.Spearman, K.J.Holzinger, S.G.Wright, J.L.Doob, R.E.Dorfman, C.H.Cramér C.R.Rao, D.A.Pierce и другие.

Различные аспекты развития и использования дельта метода рассматривается, например, в статье

A.K.Bera, M.Koley. A History of the Delta method and Some New Results. Sankhya B 85, 272–306, November (2023).

Суть этого метода заключается в следующем.
Пусть для $n \in \mathbb{N}$ имеются статистические оценки $\widehat{\theta}_n$ неизвестного параметра θ. Тогда для оценки значения $f(\theta)$, где функция f достаточно гладкая в окрестности точки θ, обычно (по методу подстановки "plug-in") используются статистики $f(\widehat{\theta}_n)$, где $n \in \mathbb{N}$. При этом, если известно предельное распределение величин $a_n(\widehat{\theta}_n - \theta)$, где $(a_n)_{n\in\mathbb{N}}$ – последовательность действительных чисел, стремящаяся к бесконечности при $n \to \infty$, то это позволяет находить предельное распределение величин $b_n(f(\widehat{\theta}_n) - f(\theta))$ для должной числовой последовательности $(b_n)_{n\in\mathbb{N}}$.

Пусть $f : \Delta \subset \mathbb{R}^N \to \mathbb{R}^k$, причем функция $f(x) = (f_1(x), \ldots, f_k(x))^\top$, где $x \in \Delta$, дифференцируема во внутренней точке $\theta \in \Delta$ (используем векторы столбцы, всюду "$\top$" обозначает транспонирование). Матрица Якоби для функции f во внутренней точке $x \in \Delta$ вводится формулой

$$J(x) = \left(\frac{\partial f_i}{\partial x_j} \right)_{i,j}, \quad i \in \{1, \ldots, k\}, \quad j \in \{1, \ldots, N\}.$$

Если $k = 1$, то $J(x)$ представляет собой транспонированный градиент $Df(x)$.

Теорема (1, E.Beutner (2023))

Пусть последовательность случайных векторов $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ *такова, что* $\widehat{\theta}_n \in \Delta$ *при* $n \in \mathbb{N}$, *и для некоторой неубывающей числовой последовательности* $(c_n)_{n\in\mathbb{N}}$, $c_n \to \infty$, *верно соотношение*

$$c_n(\widehat{\theta}_n - \theta) \xrightarrow{d} Z, \quad n \to \infty,$$

где, как обычно, "$\xrightarrow{d}$" *обозначает сходимость по распределению. Пусть для введенной выше функции* f *во внутренней точке* $\theta \in \Delta$ *определена матрица* $J(\theta)$. *Тогда справедливо соотношение*

$$c_n(f(\widehat{\theta}_n) - f(\theta)) \xrightarrow{d} J(\theta)Z, \quad n \to \infty.$$

Теорема (2, E.Beutner (2023))

Пусть для описанной выше функции f все ее частные производные вплоть до порядка r, где $r \geq 2$, непрерывны в некоторой окрестности $U(\theta) \subset \Delta$ точки $\theta \in \Delta$. Пусть последовательность случайных векторов $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ такова, что $\widehat{\theta}_n \in \Delta$ при $n \in \mathbb{N}$. Предположим, что (как в предыдущей теореме)

$$c_n(\widehat{\theta}_n - \theta) \xrightarrow{d} Z, \quad n \to \infty,$$

Тогда, если все частные производные f до порядка $r-1$ включительно обращаются в нуль в точке θ и имеется частная производная порядка r, отличная от нуля в этой точке, то

Теорема (продолжение)

$$c_n^r(f(\widehat{\theta}_n) - f(\theta)) \xrightarrow{d} W, \quad n \to \infty,$$

где компоненты вектор W *задаются формулой*

$$W_i = \frac{1}{r!} \sum_{j_1=1}^{N} \dots \sum_{j_r=1}^{N} \frac{\partial^r f(\theta)}{\partial x_{j_1} \dots \partial x_{j_r}} Z_{j_1} \dots Z_{j_r},$$

$$i = 1, \dots, k.$$

5. Предельное распределение для оценок условной информации взаимодействия

Пусть $X = (X_1, \ldots, X_m)$ – случайный вектор, заданный на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и принимающий значения в пространстве $U := U_1 \times \ldots \times U_m$, где U_i – конечное множество мощности $|U_i| = n_i$, $i = 1, \ldots, m$, $m \in \mathbb{N}$.
Положим $M = \{1, \ldots, m\}$ и $a_M = (a_1, \ldots, a_m)$, где $a_i \in U_i$, $i \in M$. Очевидно, имеется $N := n_1 \ldots n_m$ точек a_M множества U. Определим

$$p_{a_M} := \mathrm{P}(X = a_M) = \mathrm{P}(X_1 = a_1, \ldots, X_m = a_m).$$

Предположим, что дано n независимых наблюдений $X^{(1)}, \ldots, X^{(n)}$, где каждый вектор $X^{(r)}$, $r = 1, \ldots, m$, имеет такое же распределение, как вектор X. Обозначим $n_{a_M}(n)$ случайное число наблюдений среди $X^{(1)}, \ldots, X^{(n)}$, которые приняли значений a_M. Тогда N-мерный вектор с компонентами $n_{a_M}(n)$ имеет мультиномиальное распределение $Mult(n, \mathrm{p})$, где p является N-мерным вектором с компонентами p_{a_M} и $a_M \in U$. Удобно считать, что если вектор имеет компоненты, зависящие от a_M, то $(a_{i_1}, \ldots, a_{i_m})$ упорядочены с помощью лексикографического порядка для векторов $(i_1, \ldots, i_m)$.

Введем частотную оценку вероятностей p_{a_M} вида

$$\widehat{p}_{a_M}(n) := \frac{n_{a_M}(n)}{n}, \quad n \in \mathbb{N}.$$

и рассмотрим случайный вектор $\widehat{\mathbf{p}}_n$ с компонентами $\widehat{p}_{a_M}(n)$, где $a_M \in U$.
Согласно центральной предельной теореме для независимых одинаково распределенных векторов, имеющих конечные вторые моменты компонент верно соотношение

$$n^{\frac{1}{2}}(\widehat{\mathbf{p}}_n - \mathrm{p}) \xrightarrow{d} Z \sim N(0, \Sigma), \quad n \to \infty,$$

где элементы матрица Σ задаются формулой

$$\Sigma_{a_M, a'_M} = -p_{a_M} p_{a'_M} + p_{a_M} \mathbb{I}\{a_M = a'_M\}, \; a_M, a'_M \in U.$$

Напомним, что для дискретной случайной величины Z, принимающей значения $z_1, \ldots, z_k$ соответственно с вероятностями $p_1, \ldots, p_k$, энтропия Шеннона задается формулой

$$H(Z) := -\sum_{j=1}^{k} p_j \log p_j, \qquad (*)$$

где $p_j > 0$ для $j = 1, \ldots, k$ (или формально считаем $0 \log 0 := 0$). Если $X = (X_1, \ldots, X_m)$ и $T = \{i_1, \ldots, i_r\}$, то пусть $X_T := (X_{i_1}, \ldots, X_{i_v})$ и $a_T := (a_{i_1}, \ldots, a_{i_v})$. Если $X_1, \ldots, X_m$ – дискретные величины, принимающие значения в некоторых конечных множествах $U_1, \ldots, U_m$, то вводим $H(X_T)$ согласно формуле $(*)$ для $Z = X_T$.

Пусть $M := \{1, \ldots, m\}$ и $S = \{1, \ldots, m+1\}$. Нетрудно получить (как обычно, "$\subset$" обозначает нестрогое включение), что

$$I\!I(X_1, \ldots, X_m | X_{m+1}) = -\sum_{T \subset M} (-1)^{|T|} H(X_T, X_{m+1}).$$

Для рассматриваемых дискретных величин $X_1, \ldots, X_m$ можно доказать, что

$$I\!I(X_1, \ldots, X_m | X_{m+1})$$

$$= \sum_{a_S} \left(\sum_{T \subset M} (-1)^{|T|} \log(p_{a_T a_{m+1}}) \right) p_{a_S}.$$

Пусть, как и ранее, $X_k : \Omega \to U_k$, где $|U_k| = n_k$, $k \in \mathbb{N}$. Определим $N_{m+1} := n_1 \dots n_{m+1}$. Теперь (в отличие от предыдущего раздела) введем вектор $\mathbf{p}$ размерности N_{m+1} с компонентами вида p_{a_S}. Обозначим

$$f(\mathbf{p}) := II(X_1, \dots, X_m | X_{m+1}).$$

Напомним, что $X_1, \dots, X_m$ условно независимы при заданной величине X_{m+1}, если для любых $a_1, \dots, a_{m+1}$, входящих в области значений этих величин, справедливо равенство

$$\mathrm{P}(X_1 = a_1, \dots, X_m = a_m | X_{m+1} = a_{m+1})$$
$$= \prod_{k=1}^{m} \mathrm{P}(X_k = a_k | X_{m+1} = a_{m+1}).$$

Теорема (А.В.Булинский, Шаньвэнь Ван, 2024)

Пусть векторы $X_1, \ldots, X_{m+1}$ *принимают значения соответственно в конечных множествах* $U_1, \ldots, U_{m+1}$, *причем* $X_1, \ldots, X_m$ *условно независимы при заданном* X_{m+1}. *Тогда для* $f(\mathrm{p}) := I\!I(X_1, \ldots, X_m | X_{m+1})$ *верно соотношение*

$$2n(f(\widehat{\boldsymbol{p}}_n) - f(\boldsymbol{p})) \xrightarrow{d} \chi^2_K, \quad n \to \infty,$$

где $\widehat{\boldsymbol{p}}_n$ *– случайный вектор частотных оценок вектора* $\boldsymbol{p}$, *величина* χ^2_K *имеет распределение хи-квадрат с* K *степенями свободы. Здесь* $K = (n_1 - 1) \ldots, (n_m - 1)n_{m+1}$ *и* n_i *обозначает мощность множества* U_i, $i = 1, \ldots, m+1$.

Эта теорема устанавливается с помощью дельта метода. Удается явно найти гессиан H функции $f(\mathbf{p})$ и убедиться, что элементы Q_{a_S,a'_S} матрицы $Q := H\Sigma$ имеют следующий вид:

$$\mathbb{I}\{a_{m+1} = a'_{m+1}\} \sum_{T \subset M} \frac{(-1)^{m-|T|}}{p_{a_T a_{m+1}}} \mathbb{I}\{a_T = a'_T\} p_{a_T a'_{M \setminus T} a_{m+1}}.$$

Для рассматриваемых условно независимых величин доказываем идемпотентность матрицы Q, что немедленно позволяет найти ее след. Это дает ключ к нахождению предельного распределения величины $2n(f(\widehat{\mathbf{p}}_n) - f(\mathbf{p}))$ при $n \to \infty$.
Тем самым обобщается результат статьи Kubkowski M., Mielniczuk E., Teisseyre (2021), где изучался случай $m = 3$.

6. Некоторые приложения

Область FS представляет не только теоретический интерес, но и важна для разнообразных приложений.

Имеется множество взаимодополняющих средств исследований для отбора значимых факторов. Отметит различные модификации метода LASSO, использование случайных графов, MDR технику, нейронные сети, методы глубокого обучения.

Обзор B.Remeseiro and V.Bolon-Canedo (2019) посвящен FS методам в медицине.

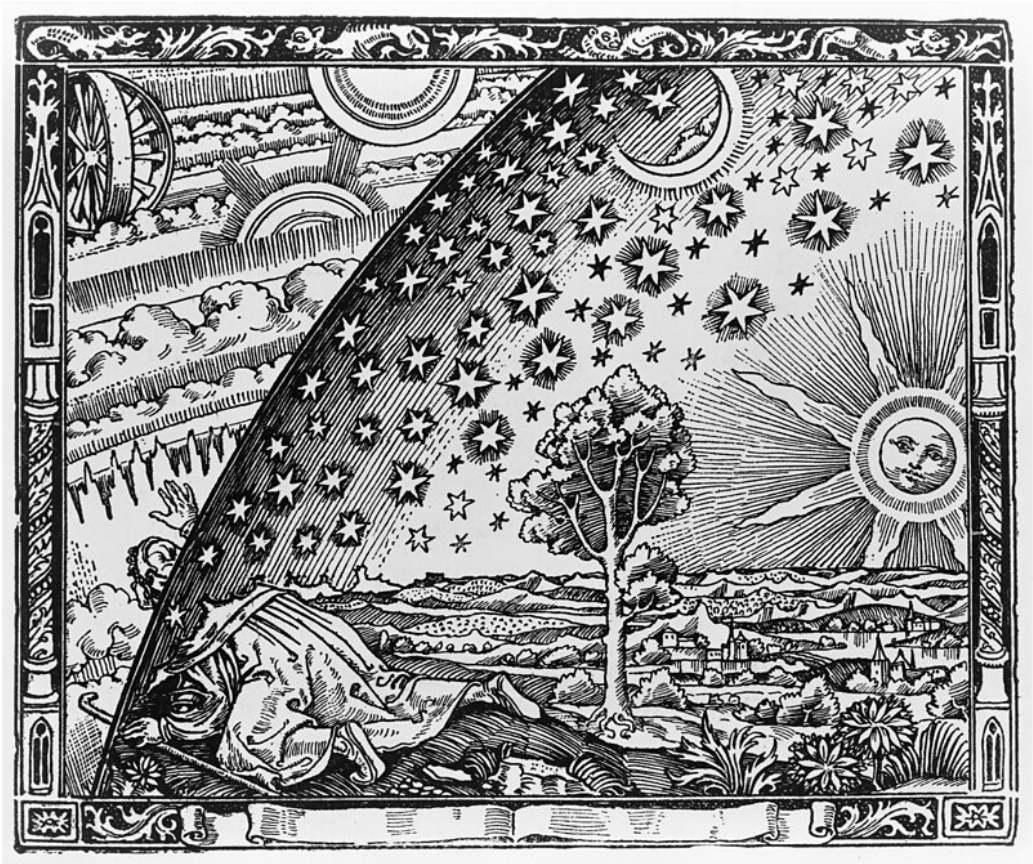
В области GWAS были успешно идентифицированы гены, которые отвечают за такие сложные заболевания как депрессивные расстройства, диабет 2-го типа, шизофрения, ишемическая болезнь сердца и другие

См., например, V.Tam et al. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews. Genetics. https://doi.org/10.1038/s41576-019-0127-1

Z.Yuan and C.Tu (2017) исследовали краткосрочные транспортные потоки, основываясь на FS с взаимной информацией.

Методы FS использованы в финансовых проблемах, см., например, B.Gültekin and B.E.Sakar (2018), X.Xia et al. (2019), K.Yu et al. (2019), A.Garcia-Medina, G.G.Farías (2020). Развитие глубокого обучения с двухступенчатой процедурой FS для анализа многомерных рядов финансовых данных рассмотрено в T.Niu et al. (2020).

7. Заключение



Гравюра из книги К.Фламмариона 1888 г.