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

Григорьева Ирина Владимировна

Канонический анализ категориальных данных с приложением в маркетинге

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Н. П. Алексеева

Рецензент:
исследователь, ВШЭ И. Б. Смирнов

Санкт-Петербург
2016

Saint Petersburg State University
Applied Mathematics and Computer Science
Computational Stochastics and Statistical Models

Grigorieva Irina Vladimirovna

Canonical analysis of categorical data with application in marketing

Bachelor's Thesis

Scientific Supervisor:
Associate Professor N. P. Alekseeva

Reviewer:
Researcher I. B. Smirnov

Saint Petersburg
2016

Содержание

Введение 5

1. Прикладная задача 5
2. Цель работы и постановка математической задачи 6
3. Исходные данные 8

Глава 1. Методы 13

1.1. Канонический анализ 13
1.2. Энтропия 14
1.3. Коэффициенты неопределенности 14
1.4. Факторный анализ 16
1.4.1. Метод главных компонент ($k \geq 2$) 17
1.5. Дисперсионный анализ 18
1.6. Алгоритм быстрого перечисления точек грассманиана 20
1.6.1. Векторная параметризация грассманиана 20
1.6.2. Отношение линейного порядка 21
1.6.3. Алгоритм быстрого перечисления точек грассманиана FGEA . . 22

Глава 2. Работа с данными 25

2.1. Множество признаков «До» и один «После» 25
2.1.1. Исследование оценок экспертов 25
2.1.2. Качество оценивания выживаемости экспертами 27
2.1.3. Изолированный анализ качественных признаков «До» 29
2.1.4. Поиск наилучшего подмножества с помощью коэффициента неопределенности 29
2.2. Множества признаков «До» и «После» 35
2.2.1. Перебор подмножеств и поиск наиболее связанных с помощью коэффициента неопределенности 35
2.2.2. Частотный способ поиска номинативных представителей 36
2.2.3. Метод поиска номинативных представителей, основанный на удалении признаков 45
2.2.4. Факторный анализ для поиска номинативных представителей . . 54

2.3. Заключение 57

Литература 60

Введение

В данной работе рассматривается задача исследования зависимости между двумя множествами признаков, а именно: между различными комбинациями начальных и итоговых признаков базы данных, полученной от исследователя. Производится поиск связей между множествами, где в качестве меры зависимости рассматривается коэффициент неопределенности. Для расширения исходных множеств используются алгебраические методы: алгоритм быстрого перечисления точек грассманиана. Трудности анализа: сложная итоговая характеристика, задачу нельзя решить напрямую, интерпретация полученных связей. Отличие моей работы от других: поиск не только наибольших связей между множествами, но и самых устойчивых симптомов, в смысле уменьшения количества значимых связей и снижения уровней зависимости при их исключении из совокупностей, названных номинативными представителями. Задача нахождения «сильных» связей между наборами признаков важна в реальной жизни, потому что на основе полученных результатов принимаются решения в той или иной сфере жизни. Например, ежегодно проходит большое количество конкурсов для молодых ученых и жюри не должно ошибаться в выборе победителей.

1. Прикладная задача

Прошел конкурс «Инновации в Образовании», на который в 2014 году прислали много заявок.

Были получены данные, которые состоят из трех блоков:

- Первый блок — это информация из 552 заявки на конкурс «Инновации в образовании».
- Второй блок — это оценки экспертов к каждой из заявок. Эксперт мог отметить, что заявка «бракованная» и не выставлять подробных оценок, в противном случае — оценивал по нескольким критериям.
- Третий блок — это анкета, которую участники заполняли через год. На нее ответило 240 человек. Они могли указать, продолжают ли работу над проектом или уже забросили. Те, кто продолжают, отвечали на ряд вопросов.

Возникает прикладная задача:

Найти признаки«До», которые оказывают самое сильное влияние на дальнейшее развитие проекта.

2. Цель работы и постановка математической задачи

Целью работы является исследование зависимости между двумя наборами качественных признаков.

Номинальные признаки представлены категориями, для которых не определен никакой другой способ сравнения, кроме как буквальное совпадение или несовпадение.

Имеется набор итоговых характеристик и исходный набор признаков. База данных (заявки на участие в конкурсе) разделена на период «До» и «После» и выбраны только номинальные признаки.

Введем несколько необходимых определений [1]:

Симптомом ранга k называется $\forall$ линейная комбинация вида $X_\tau = A_\tau X (\mathrm{mod}\, 2)$, где
вектор $X = (X_1, \ldots, X_m)$ с компонентами, принимающими значения 1 и 0,
$\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$ k-подмножество из m натуральных чисел,
вектор-строка $A_\tau = (a_1, \ldots, a_m)$ с компонентами

$$a_j = \begin{cases} 1, & j \in \tau \\ 0, & \text{иначе} \end{cases}$$

Компоненты вектора X являются тривиальными симптомами единичного ранга $X_i, i = 1, \ldots, m$. Симптом нулевого ранга, то есть со всеми нулевыми коэффициентами, является вырожденным $X_\emptyset$ и принимает значение 0 с вероятностью 1.
Симптом — это новый признак, отражающий взаимодействие исходных признаков. Он может нести в себе информацию, не содержащуюся в исходных признаках по отдельности. Таким образом, симптомы позволяют исследовать взаимодействие бинарных признаков без увеличения размерности.

Пусть имеется $k+1 \geq 0$ симптомов $X_0, \ldots, X_k$.

Синдромом k-го порядка называется совокупность $2^{k+1}-1$ симптомов вида

$$\beta_1 X_0 + \ldots + \beta_k X_k (\mathrm{mod}\, 2),$$

где $\beta_i \in \mathbb{F}_2$ не равны нулю одновременно.

Номинативный представитель симптом наименьшего ранга, без которого нельзя получить значимые связи между множествами признаков.

Математическая задача: Поиск подмножеств признаков «До» и «После», связанных наилучшим образом, и номинативных представителей этих множеств.

Используемые методы:

1. Канонический анализ.
2. Коэффициент неопределенности.
3. Факторный анализ.
   - Исследование оценок экспертов.
   - Редукция размерности.
4. Дисперсионный анализ.
   - Качество оценивания выживаемости экспертами.
5. Алгоритм быстрого перечисления точек грассманиана Ананьевской П.В [2].

3. Исходные данные

Таблица 1. Признаки «До» (X).

OS	Операционная система с которой подавалась заявка, 1 — Windows, 2 — Mac OS, 3 — Linux.
$LOCATION$	Город из которого подавалась заявка, 1 — Москва, 2 — Санкт-Петербург, 3 — другой российский город, 4 — город СНГ, 5 — другой город.
$KINDERGARTEN$	Входит ли дошкольное образование в сферу проекта, 0 — нет, 1 — да.
$PRIMARY\_SCHOOL$	Входит ли начальное образование в сферу проекта, 0 — нет, 1 — да.
$MIDDLE\_SCHOOL$	Входит ли среднее образование в сферу проекта, 0 — нет, 1 — да.
$HIGH\_SCHOOL$	Входит ли старшая школа в сферу проекта, 0 — нет, 1 — да.
$UNIVERSITY$	Входит ли высшее образование в сферу проекта, 0 — нет, 1 — да.
$EXTRACURRICULAR$	Входит ли дополнительное образование в сферу проекта, 0 — нет, 1 — да.
$PROFESSIONAL$	Входит ли профессиональное образование в сферу проекта, 0 — нет, 1 — да.
$FAMILY$	Входит ли семейное образование в сферу проекта, 0 — нет, 1 — да.
$OTHER$	Входит ли иное образование в сферу проекта, 0 — нет, 1 — да.
WEB_SITE	Наличие веб-сайта,0 — нет, 1 — да.
SEX	Половой состав команды, 1 — только мужчины, 2 — только женщины, 3 — смешанный состав.
$TEACHER$	Есть ли в команде преподаватель, 0 — нет, 1 — да.
$ENTREPRENEUR$	Есть ли в команде предприниматель, 0 — нет, 1 — да.
$INDUSTRY$	Есть ли в команде сотрудник организации, 0 — нет, 1 — да.

В таблице 1 расшифровка наименований первого множества — признаки «До».

К признакам «До» относится оценка успешности проекта экспертами — *JurySection* (табл. 2).

Таблица 2. *JurySection*.

JURY	Эксперт, Разные числа соответствуют разным экспертам.
JURY_OVERALL	Общая оценка экспертом, 0 - далее не рассматривать, 1 — рассматривать в общем порядке, 2 — обратить особое внимание.
JURY_NOVELTY	Новизна и оригинальность идеи, От 1 до 5, 5 — лучше всего.
JURY_IMPORTANCE	Актуальность решаемых проблем, От 1 до 5, 5 — лучше всего.
JURY_RELEVANCE	Целесообразность используемых механизмов, От 1 до 5, 5 — лучше всего.
JURY_SCALABILITY	Возможность тиражирования, От 1 до 5, 5 — лучше всего.

Второе множество признаков — признаки «После» (Y) (табл. 3, табл. 4).

Таблица 3. Признаки «После» (Y).

$EVENTS\_SITE$	Был запущен сайт, 0 — нет, 1 — да.
$EVENTS\_PUBLICATION$	Публикации в СМИ о проекте, 0 — нет, 1 — да.
$EVENTS\_STAFF$	Наняты новые сотрудники, 0 — нет, 1 — да.
$EVENTS\_PILOT$	Прошел запуск пилота проекта, 0 — нет, 1 — да.
$EVENTS\_PRODUCTION$	Запущено производство, 0 — нет, 1 — да.
$EVENTS\_BRANCH$	Открылось новое отделение/представительство, 0 — нет, 1 — да.
$EVENTS\_PARTNERS$	Привлечены новые партнеры, 0 — нет, 1 — да.
$EVENTS\_SUPPORT$	Получена административная поддержка, 0 — нет, 1 — да.
$EVENTS\_INCOME$	Увеличилась выручка, 0 — нет, 1 — да.
$EVENTS\_GRANT$	Получен грант, 0 — нет, 1 — да.
$EVENTS\_INVESTMENT$	Привлечены инвестиции, 0 — нет, 1 — да.

Таблица 4. Признаки «После» (Y): Насколько команда была вовлечена в следующие активности.

ACTIVITIES_EVENTS	Посещение тематических мероприятий, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_ONLINE	Прохождение онлайн-курсов, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_LITERATURE	Чтение специальной литературы, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_RESEARCH	Поиск исследований, подтверждающих потенциал проекта, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_COMPETITORS	Поиск аналогичных проектов, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_MENTORS	Общение с экспертами и менторами, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
ACTIVITIES_COMPETITIONS	Участие с проектом в конкурсах, 0 — никогда, 1 — однократно, 2 — несколько раз, 3 — регулярно.
TEAM_SEX	Пол ключевых членов команды, 1 — все мужчины, 2 — все женщины, 3 — мужчины и женщины.
TEAM_TOP_MSC	Количество ключевых участников команды закончивших ведущий вуз Москвы, 0 — ни одного, 1 — один, 2 — больше одного.
TEAM_TOP_SPB	Количество ключевых участников команды закончивших ведущий вуз Санкт-Петербурга, 0 — ни одного, 1 — один, 2 — больше одного.

$TEAM\_TOP\_PROVINCE$	Количество ключевых участников команды закончивших ведущий вуз не Москвы и Петербурга, 0 — ни одного, 1 — один, 2 — больше одного.
$FOREIGN$	Количество ключевых участников команды имеющих заграничных опыт, 0 — нет, 1 — да.
$IS\_EDUCATION$	Есть ли среди ключевых участников команды специалист в области образование и педагогика, 0 — нет, 1 — да.
$IS\_ECONOMICS$	Есть ли среди ключевых участников команды специалист в области экономика и управление, 0 — нет, 1 — да.
IS_MATH	Есть ли среди ключевых участников команды специалист в области математика, программирование, технические науки, 0 — нет, 1 — да.
$IS\_HUMANITIES$	Есть ли среди ключевых участников команды специалист в области гуманитарные и социальные науки, 0 — нет, 1 — да.
$IS\_NATURAL$	Есть ли среди ключевых участников команды специалист в области естественные науки, 0 — нет, 1 — да.
$IS\_CULTURE$	Есть ли среди ключевых участников команды специалист в области культура и искусство, 0 — нет, 1 — да.

Глава 1

Методы

1.1. Канонический анализ

Канонический анализ позволяет определить взаимосвязь между двумя совокупностями признаков, характеризующих объекты [3]. Например, можно изучить зависимость между различными неблагоприятными факторами и появлением определенной группы симптомов заболевания, или взаимосвязь между двумя группами синдромов больного.

Корреляция это степень зависимости между $\xi = (\xi_1, \ldots, \xi_n)$ и $\eta = (\eta_1, \ldots, \eta_n)$. Она выражается через коэффициент корреляции

$\mathrm{R}(\xi, \eta) = \dfrac{\sum_{i=1}^{n} (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})}{\sqrt{\sum_{i=1}^{n} (\xi_i - \bar{\xi})^2} \sqrt{\sum_{i=1}^{n} (\eta_i - \bar{\eta})^2}}$, всегда $|\,\mathrm{R}\,| \leq 1$.

Если признаки независимые, то $\mathrm{R} = 0$.

Обычные коэффициенты корреляции используются для выявления линейной зависимости между двумя признаками X и Y. Если нужно выявить зависимость между X_0 и $X_1, \ldots, X_p$, то в качестве характеристики этой зависимости рассматривается множественный коэффициент корреляции, равный коэффициенту корреляции $\mathrm{R}(X_0, \hat{X_0})$, где $\hat{X_0} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ наилучшее линейное предсказание X_0.

Эта концепция была обобщена на случай связи между множествами признаков, характеризующих объекты.

Канонический анализ является обобщением множественной корреляции как меры связи между одной переменной и множеством других переменных. [4]

Задача здесь состоит в том, чтобы найти такие нормированные линейные комбинации:

$$U_1 = \beta_{10} + \beta_{11} X_1 + \ldots + \beta_{1r} X_r,$$

$$V_1 = \alpha_{10} + \alpha_{11} X_{r+1} + \ldots + \alpha_{1s} X_{r+s},$$

таким образом, чтобы каноническая корреляция $\mathrm{R} = \mathrm{cor}(U_1, V_1)$ была максимальной (т.е надо найти весовые коэффициенты таким образом, чтобы каноническая корреляция была максимальной).

Проблема:

1. Выбор метрики.
2. Перебор синдромов.
3. Поиск синдромов признаков «До» и «После», связанных наилучшим образом.

Метрика величина, которая измеряет связь между двумя наборами признаков.

В качестве метрики используется коэффициент неопределенности.

1.2. Энтропия

Пусть задана случайная величина $\xi = \begin{pmatrix} x_1 & \cdots & x_k \\ p_1 & \cdots & p_k \end{pmatrix}$

Энтропия $\mathrm{H}(\xi) = -\sum_{i=1}^{k} p_i \log_2 p_i$.

Наименьшее значение энтропия принимает, когда случайная величина постоянна. Если $\xi = c$, то $\mathrm{H}(\xi) = 0$ – неопределенности нет. Наибольшее значение H принимает в случае, когда ξ имеет равномерное распределение , т. е $p_i = \frac{1}{k}$: $\mathrm{H}(\xi) = \log_2 k$.

Энтропия характеризует степень неопределенности и является информационной характеристикой случайной величины.

1.3. Коэффициенты неопределенности

Пусть задан набор из l дискретных случайных векторов

$X^{(s)} = (X_1{}^{(s)}, \ldots, X_{m_s}{}^{(s)}), s = 1, \ldots, l.$

Мерой зависимости двух случайных векторов может быть выбран односторонний или двусторонний коэффициент неопределенности Тейла [4]:

Односторонний коэффициент неопределенности между двумя векторами $X^{(r)}$ и $X^{(s)}$ вычисляется по формуле

$$\mathrm{J}_0(X^{(r)}|X^{(s)}) = \frac{(\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)}) - \mathrm{H}(X^{(r)}, X^{(s)}))100\%}{\mathrm{H}(X^{(s)})}, \tag{1.1}$$

$$\mathrm{J}_0(X^{(s)}|X^{(r)}) = \frac{(\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)}) - \mathrm{H}(X^{(r)}, X^{(s)}))100\%}{\mathrm{H}(X^{(r)})}, \tag{1.2}$$

где $\mathrm{H}(X^{(r)}, X^{(s)})$ — энтропия вектора $({X_1}^{(r)}, \ldots, {X_{m_r}}^{(r)}, {X_1}^{(s)}, \ldots, {X_{m_s}}^{(s)})$.

Энтропия распределения случайного вектора $X = (X_1 ... X_m)^{\mathrm{T}}$ вычисляется по формуле:

$$\mathrm{H}(X) = -\sum_{i=1}^{q^m} p_i \log_2 p_i,$$

где $p_i = \mathrm{P}(X = (q_1, \ldots, q_m))$.

Двусторонний коэффициент неопределенности может быть задан следующим образом:

$$\mathrm{J}(X^{(r)}, X^{(s)}) = 2\frac{\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)}) - \mathrm{H}(X^{(r)}, X^{(s)})}{\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)})}. \tag{1.3}$$

Заметим, что односторонний и двусторонний коэффициенты неопределенности представляют собой нормализованные версии совместной информации $\mathrm{I}(X^{(r)}, X^{(s)}) = \mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)}) - \mathrm{H}(X^{(r)}, X^{(s)})$, являющейся, в свою очередь, одной из наиболее известных мер независимости.

Запишем коэффициенты (1.1), (1.2) и (1.3) таким образом:

$$\mathrm{J}_0(X^{(r)}|X^{(s)}) = \frac{\mathrm{I}(X^{(r)}, X^{(s)})}{\mathrm{H}(X^{(s)})} 100\%,$$

$$\mathrm{J}_0(X^{(s)}|X^{(r)}) = \frac{\mathrm{I}(X^{(r)}, X^{(s)})}{\mathrm{H}(X^{(r)})} 100\%,$$

$$\mathrm{J}(X^{(r)}, X^{(s)}) = 2\frac{\mathrm{I}(X^{(r)}, X^{(s)})}{\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)})} =$$

$$= \frac{\mathrm{H}(X^{(r)})}{\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)})} \mathrm{J}_0(X^{(s)}|X^{(r)}) + \frac{\mathrm{H}(X^{(s)})}{\mathrm{H}(X^{(r)}) + \mathrm{H}(X^{(s)})} \mathrm{J}_0(X^{(r)}|X^{(s)}).$$

Статистика $\mathrm{J}(X^{(r)}, X^{(s)})$ является симметричной и измеряет количество информации в переменной $X^{(r)}$ относительно переменной $X^{(s)}$ или в переменной $X^{(s)}$ относительно переменной $X^{(r)}$. Статистики $\mathrm{J}_0(X^{(r)}|X^{(s)})$ и $\mathrm{J}_0(X^{(s)}|X^{(r)})$ выражают направленную зависимость: показывают, сколько информации об $X^{(s)}$ дает знание $X^{(r)}$ и наоборот.

Значение совместной информации и коэффициентов неопределенности достигает нуля в случае независимости $X^{(r)}$ и $X^{(s)}$

Было доказано утверждение:

Утверждение 1. *Пусть* x_{τ_i} *— симптомы* $i = 1, \ldots, n$, $\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle$, $\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle$ *и* y *— синдромы. Односторонние коэффициенты неопределенности:*

$$\mathrm{J}_1 = \mathrm{J}_0(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle|y) = \frac{\mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)}{\mathrm{H}(y)},$$

$$\tilde{\mathrm{J}}_1 = \mathrm{J}_0(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle|y) = \frac{\mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)}{\mathrm{H}(y)}.$$

1. *Если симптомы* $x_{\tau_1}, \ldots, x_{\tau_n}$ *независимы, то* $\mathrm{J}_1 - \tilde{\mathrm{J}}_1 > 0$,

2. *Если симптомы* $x_{\tau_1}, \ldots, x_{\tau_n}$ *и* y *независимы, то* $\mathrm{J}_1 - \tilde{\mathrm{J}}_1 = 0$.

Доказательство:

$\mathrm{H}(y)(\mathrm{J}_1 - \tilde{\mathrm{J}}_1) = \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) - (\mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)) = -(\mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle)) + (\mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)) =$

Прибавим и отнимем $\mathrm{H}(x_{\tau_1})$:

$= -(\mathrm{H}(x_{\tau_1}) + \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle)) + (\mathrm{H}(x_{\tau_1}) + \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) - \mathrm{H}(\langle x_{\tau_1}, x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)) = -\mathrm{I}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{I}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y)$, где I —совместная информация.

1) Если $x_{\tau_1}, \ldots, x_{\tau_n}$ независимы $\Rightarrow \mathrm{I}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) = 0$.

2) Если $x_{\tau_1}, \ldots, x_{\tau_n}$ и y независимы $\Rightarrow \mathrm{I}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) = 0$ и $\mathrm{I}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) = \mathrm{H}(x_{\tau_1}) + \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) - \mathrm{H}(x_{\tau_1}, y) - \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) + \mathrm{H}(x_{\tau_1}, \langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle, y) = \mathrm{H}(x_{\tau_1}) + \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) - \mathrm{H}(x_{\tau_1}) - \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) - \mathrm{H}(x_{\tau_1}) - \mathrm{H}(y) - \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) - \mathrm{H}(y) + \mathrm{H}(x_{\tau_1}) + \mathrm{H}(\langle x_{\tau_2}, \ldots, x_{\tau_n}\rangle) + \mathrm{H}(y) = 0$, что и.т.д.

1.4. Факторный анализ

Задачей факторного анализа является объединение большого количества признаков, которыми характеризуется объект, в меньшее количество искусственно построенных на их основе факторов, чтобы полученная в итоге система факторов была наиболее удобна с точки зрения содержательной интерпретации.[4]

Методы факторного анализа различают в зависимости от подходов для нахождения коэффициентов значения факторов. В работе использовался метод главных компонент. Он основан на определении минимального числа факторов, которые вносят наибольший вклад в дисперсию данных. Они называются главными компонентами.

1.4.1. Метод главных компонент ($k \geq 2$)

Идея: Заменить k-ую случайную величину при наименьшей потери информативности на m-ую ($m < k$).

Эффективность любого выбора зависит от того, в какой степени эти m линейных функций дают возможность реконструировать или восстановить k первоначальных величин. Один из методов реконструкции этой первоначальной случайной величины состоит в построении ее наилучшего предиктора на основе m линейных функций.

Наилучший выбор линейных функций: первые m главных компонент.

Пусть $X_1, ..., X_k$ — признаки.

Первой главной компонентой Y_1 называется сохраняющая расстояние между точками линейная комбинация исходных признаков

$$Y_1 = \alpha_{11}X_1 + ... + \alpha_{k1}X_k,$$

где коэффициенты $\alpha_{11}, ..., \alpha_{k1}$ выбираются таким образом, чтобы дисперсия $\mathrm{D}(Y_{11}) = \lambda_1$ была максимальной, т.е по Y_1 индивиды отличаются наибольшим образом.

Вторая главная компонента также является линейной комбинацией исходных признаков:

$$Y_2 = \alpha_{12}X_1 + ... + \alpha_{k2}X_k,$$

где коэффициенты $\alpha_{12}, ..., \alpha_{k2}$ выбираются таким образом,что компоненты Y_1 и Y_2 некоррелированы, а дисперсия $\mathrm{D}(Y_2) = \lambda_2$ является максимальной из всех линейных комбинаций, некоррелированных с Y_1, то есть вторая компонента должна нести наибольшую новую информацию, не имеющую отношения к первой главной компоненте. Аналогично строятся остальные главные компоненты:

$$Y_j = \sum_{i=1}^{k} \alpha_{ij}X_i, \ j = 1..k. \tag{1.4}$$

Суммарная дисперсия остается неизменной:

$$V = \mathrm{D}(X_1) + ... + \mathrm{D}(X_k) = \lambda_1 + ... + \lambda_k.$$

Значимость главных компонент (1.4) определяется долей объясняемой ими дисперсии, равной $\frac{\lambda_i}{V}100\%$.

Факторами называются нормированные главные компоненты $\frac{Y_i}{\sqrt{\lambda_i}}$.

1.5. Дисперсионный анализ

Задачей дисперсионного анализа является изучение влияния одного или нескольких факторов на рассматриваемый признак.

Целью дисперсионного анализа является проверка значимости различия между средними в разных группах с помощью сравнения дисперсий этих групп. Разделение общей дисперсии на несколько источников, позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью.

Однофакторный дисперсионный анализ используется в тех случаях, когда в распоряжении имеется выборка, которая разбивается на r групп.[5]
Требуется проверить гипотезу о равенстве средних:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_r$

H_1 : не все средние равны.

При истинности нулевой гипотезы, оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии. При ложности — значимо отклоняться.

Для проверки этой гипотезы используется критерий Фишера.
Пусть x_{ik} — i-ый элемент $(i = 1...n_k)$ k выборки, где n_k — число данных в k выборке. Тогда $\bar{x_k}$ — выборочное среднее k — выборки определяется по формуле

$\bar{x_k} = \dfrac{1}{n_k} \sum_{i=1}^{n_k} x_{ik}.$

Общее среднее вычисляется по формуле $\bar{x} = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}$, где $n = \sum_{k=1}^{m} n_k$.

Основное тождество дисперсионного анализа имеет следующий вид: $Q = Q_1 + Q_2$, где

Q_1 — сумма квадратов отклонений выборочных средних $\bar{x_k}$ от общего среднего $\bar{x}$ (сумма квадратов отклонений между группами);

Q_2 — сумма квадратов отклонений наблюдаемых значений x_{ik} от выборочной средней $\bar{x_k}$ (сумма квадратов отклонений внутри групп);
Q — общая сумма квадратов отклонений наблюдаемых значений x_{ik} от общего среднего $\bar{x}$.

Расчет этих сумм квадратов отклонений осуществляется по следующим формулам:

$$Q = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik})^2 - n(\bar{x})^2,$$

$$Q_1 = \sum_{k=1}^{m} (\bar{x_k} - \bar{x})^2 = \sum_{k=1}^{m} n_k(\bar{x_k})^2 - n(\bar{x})^2,$$

$$Q_2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \bar{x_k})^2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik})^2 - \sum_{k=1}^{m} n(\bar{x_k})^2.$$

В качестве критерия необходимо воспользоваться критерием Фишера:

$$F = \frac{Q_1/(m-1)}{Q_2/(n-m)}.$$

Если расчетное значение критерия Фишера будет меньше, чем табличное значение $F_{\lambda;m-1;n-m}$ — нет оснований считать, что независимый фактор оказывает влияние на разброс средних значений, в противном случае, независимый фактор оказывает существенное влияние на разброс средних значений (λ — уровень значимости, обычно для задач $\lambda = 0.05$).

1.6. Алгоритм быстрого перечисления точек грассманиана

1.6.1. Векторная параметризация грассманиана

Грассманиан (многообразие Грассмана) совокупность всевозможных гиперпространств фиксированной размерности некоторого векторного пространства над произвольным полем.

Векторная параметризация грассманиана над конечным полем является модификацией классического клеточного разложения и позволяет решить задачу быстрого перечисления точек этого многообразия.[2]

Линейно независимые вектора $X_1, \ldots, X_m$ над конечным полем $\mathbb{F}_q$ — базис m-мерного пространства $V_m = (\mathbb{F}_q)^m$, т.е $\forall X_\tau \in V_m$: $X_\tau = a_1X_1 + \ldots + a_mX_m$, где $a_i \in \mathbb{F}_q, \tau = \{a_i\}_{a_i \neq 0}$.
Рассмотрим набор линейно независимых векторов $(X_{\tau_1}, \ldots, X_{\tau_k})$ как базис, образующий k-мерное подпространство V_k пространства V_m. Всевозможные k-мерные подпространства образуют грассманиан $\mathrm{Gr}_q(k, m)$, точкой которого является одно k-мерное подпространство.
Зафиксируем полный флаг $\mathcal{F}$ на пространстве V_m:

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

Введем несколько определений, чтобы сформулировать теорему, на которой будет основываться алгоритм:

Отношение линейного порядка Бинарное отношение $v \prec w$ на множестве векторов пространства V_m:

1. рефлексивность: $\forall v \in V_m \; v \prec v$;
2. транзитивность: $\forall u, v, w \in V_m \; u \prec v$ и $v \prec w \Rightarrow u \prec w$;
3. антисимметричность: $\forall v, w \in V_m \; v \prec w$ и $w \prec v \Rightarrow v = w$;
4. $\forall v, w \in V_m$ если $v \prec w$ или $w \prec v$.

Отношение линейного порядка $v \prec w$ на V_m согласовано с флагом $\mathcal{F}$, если для $\forall i \; v \in V_i$, $w \in V_m \backslash V_i \Rightarrow v \prec w$.

Зададим такую векторную параметризацию грассманиана $\mathrm{Gr}_q(k,m)$, с помощью которой будет удобно перечислять всевозможные его точки (единообразно выделим единственный базис $(X_{\tau_1},\ldots,X_{\tau_k})$ в каждом k-мерном подпространстве).

Теорема 1 (о векторной параметризации, Ананьевская П. В.) *Для пространства V_m, полного флага $\mathcal{F}$ и согласованного с ним отношения линейного порядка отображение*

$$(X_{\tau_1},\ldots,X_{\tau_k}) \longmapsto \langle X_{\tau_1},\ldots,X_{\tau_k}\rangle$$

устанавливает биекцию между наборами векторов $X_{\tau_1},\ldots,X_{\tau_k} \in V_m$ такими, что

1. *для $X_{\tau_i} = a_{i_1}X_1 + \ldots + a_{i_m}X_m$ имеет место $(a_{i_1},\ldots,a_{i_m}) = (a_{i_1},\ldots,a_{i_s},1,0,\ldots,0)$, где $s \le m$.*
2. *$X_{\tau_i} \prec X_{\tau_j}$ при $i \le j$,*
3. *для всех X_{τ_j}, $i \le j$ выполнено $a_{j(s_i+1)} = 0$,*

и k- мерными подпространствами V_m.

1.6.2. Отношение линейного порядка

Зададим отношение линейного порядка на множестве векторов пространства V_m. Основное условие на это отношение: согласованность с флагом $\mathcal{F}$.

Лексикографическим порядком $\succeq_l$ на $\mathbb{F}_q^m$ называется отношение линейного порядка, если для $\forall\ (a_1,\ldots,a_m),(a'_1,\ldots,a'_m) \in \mathbb{F}_q^m$ верно правило

$$(a_1,\ldots,a_m) \succeq_l (a'_1,\ldots,a'_m) \Longleftrightarrow \sum_{i=1}^{m} a_i q^{i-1} \ge \sum_{i=1}^{m} a'_i q^{i-1}.$$

Обобщенным порядком Грея $\succeq_g$ называется отношение линейного порядка, если $(a_1,\ldots,a_m) \succeq_g (a'_1,\ldots,a'_m)$ тогда и только тогда, когда

$$(a_1 \oplus a_2 \oplus \ldots \oplus a_m, \ldots, a_{m-1} \oplus a_m, a_m) \succeq_l (a'_1 \oplus a'_2 \oplus \ldots \oplus a'_m, \ldots, a'_{m-1} \oplus a'_m, a'_m),$$

где суммы считаются по модулю q.

Лексикографический порядок и обобщенный порядок Грея согласованы с флагом $\mathcal{F}$. Преимущество обобщенного порядка Грея состоит в том, что каждая следующая строка $(a_1,\ldots,a_m)$ отличается от предыдущей $(a'_1,\ldots,a'_m)$ прибавлением 1 (по модулю q) ровно к одному из a'_i.

1.6.3. Алгоритм быстрого перечисления точек грассманиана FGEA

Алгоритм основан на векторной параметризации грассманиана и ориентирован на сокращение количества операций для построения каждой следующей точки за счет использования обобщенного кода Грея и соответствующего ему отношения линейного порядка.

Для того, чтобы перечислить все точки грассманиана $\mathrm{Gr}_q(k,m)$, т.е. все возможные k-мерные векторные подпространства пространства V_m, достаточно перебрать базисы этих подпространств (всевозможные наборы $X_{\tau_1},\ldots,X_{\tau_k}$). Однако при таком подходе все подпространства будут учтены по нескольку раз, например наборы $(X_1,X_2,\ldots,X_k)$ и $(X_1+X_2,X_2,\ldots,X_k)$ задают одно и то же подпространство. Поэтому требуется описать некоторую процедуру, позволяющую избежать повторений такого рода.

Все векторы X_{τ_i} являются линейными комбинациями линейно независимых векторов $X_1,\ldots,X_m$, существуют единственные наборы коэффициентов $a_{i1},\ldots,a_{im}\in\mathbb{F}_q$:

$$X_{\tau_i}=a_{i1}X_1+\ldots+a_{im}X_m.$$

Таким образом, есть взаимно однозначное соответствие между наборами векторов $X_{\tau_1},\ldots,X_{\tau_k}$ и матрицами коэффициентов $A=\{a_{ij}\}_{i,j}^{k,m}$.

Согласно теореме о векторной параметризации, для того, чтобы перечислить все точки грассманиана ровно по одному разу, достаточно рассматривать только наборы коэффициентов (a_{ij}), обладающие следующими свойствами:

1. для $\forall\ i$ найдется индекс s_i: $(a_{i1},\ldots,a_{im})=(a_{i1},\ldots,a_{is_i},1,0,\ldots,0)$.
2. $a_{ij}=0$, если найдется индекс l: $l>i$ и $s_l<j$.
3. $a_{i(s_j+1)}=0$ для $i\neq j$.

Другими словами, матрица A должна иметь вид:

$$\begin{array}{c} \\ X_{\tau_1} \\ X_{\tau_2} \\ \vdots \\ X_{\tau_i} \\ \vdots \\ X_{\tau_k} \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccccccc} X_1 & \ldots & & & & \ldots & & & & \ldots & & X_s & & \ldots & & & & \ldots & X_m \end{array} \\
\left[\begin{array}{cccccccccccccccccccc}
* & \ldots & * & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
* & \ldots & * & 0 & * & \ldots & * & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & & & & & \vdots \\
* & \ldots & * & 0 & * & \ldots & * & 0 & * & \ldots & * & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & & & & & \vdots \\
* & \ldots & * & 0 & * & \ldots & * & 0 & * & \ldots & * & 0 & * & \ldots & * & 1 & 0 & \ldots & 0
\end{array}\right]
\end{array}$$

Следовательно, для эффективного перечисления точек грассманиана достаточно уметь перебирать все матрицы A указанного вида. Для этого будут последовательно формировать вектора X_{τ_i}, т.е. строки матрицы A.

С вычислительной точки зрения, для перебора векторов X_{τ_i} наиболее эффективно использовать упорядочивание, соответствующее обобщенному порядку Грея.

Алгоритм представлен в следующем виде:

Цикл 1 Для заданного набора $X^{(1)} = (X_1, \ldots, X_m)$:

а) в порядке кодирования Грея перебираем все наборы $(a_{11}, \ldots, a_{1m})$, формируя последовательно на каждой итерации цикла

$$(X_{\tau_1})_{iter_1} = (X_{\tau_1})_{iter_1 - 1} + a_{1t} X_t^{(1)},$$

где a_{1t} – единственный элемент отличающий текущий набор от предыдущего, а $X_t^{(1)} = X_t$.

б) для текущего вектора $(X_{\tau_1})_{iter_1}$ и соответствующего ему набора коэффициентов $(a_{11}, \ldots, a_{1s_1}, 1, 0, \ldots, 0)$ определяем максимальный номер $j_1 = s_1 + 1 : a_{1(j_1+1)} = 0$.

Цикл 2 Вычеркиваем из набора $X^{(1)}$ вектор X_{j_1} и для набора

$$X_{(2)} = (X_1, \ldots, X_{j_1 - 1}, X_{j_1 + 1}, \ldots, X_m) :$$

а) в порядке кодирования Грея перебираем все наборы $(a_{21}, \ldots, a_{2(m-1)})$, начиная с набора $(0, \ldots, 0, 1, 0, \ldots, 0)$, где единица стоит на j_1-месте, и формируя последовательно на каждой итерации

$$(X_{\tau_2})_{iter_2} = (X_{\tau_2})_{iter_2 - 1} + a_{2t} X_t^{(2)},$$

где $X_t^{(2)}$ – вектор, стоящий на месте t в наборе $X^{(2)}$.

б) для текущего вектора $(X_{\tau_2})_{iter_2}$ и соответствующего ему набора коэффициентов $(a_{21}, \ldots, a_{2s_2}, 1, 0, \ldots, 0)$ определяем максимальный номер $j_2 = s_2 + 1 : a_{2(j_2+1)} = 0$.

...

Цикл k Вычеркиваем из набора $X^{(1)}$ вектора $X_{j_1}, X_{j_2}, \ldots, X_{j_{(k-1)}}$ и для набора

$$X_{(k)} = (X_1, \ldots, \widehat{X}_{j_1}, \ldots, \widehat{X}_{j_{(k-1)}}, \ldots, X_m) :$$

а) в порядке кодирования Грея перебираем все наборы $(a_{k1}, \ldots, a_{k(m-(k-2))})$, начиная с набора $(0, \ldots, 0, 1, 0, \ldots, 0)$, где единица стоит на $j_{(k-1)} + 2 - k$-месте, и формируя последовательно

$$(X_{\tau_k})_{iter_k} = (X_{\tau_k})_{iter_k - 1} + a_{kt} X_t^{(k)},$$

где $X_t^{(k)}$ – вектор, стоящий на месте t в наборе $X^{(k)}$.

б) Составляем базис подпространства из текущих векторов:

$$V_k^{(iter)} = \langle (X_{\tau_1})_{iter_1}, \ldots, (X_{\tau_k})_{iter_k} \rangle$$

конец k-го цикла

конец 2-го цикла

конец 1-го цикла

Глава 2

Работа с данными

2.1. Множество признаков «До» и один «После»

2.1.1. Исследование оценок экспертов

Рассмотрена *JurySection* — секция, которая содержит оценки экспертов для каждого проекта. Оценка ставилась экспертом после ознакомления с анкетой проекта. Оценок – признаков достаточно много, хотелось бы уменьшить количество переменных, обобщив их, используя факторный анализ.

Был проведен факторный анализ в *Statistica* 7 для четырех признаков $JURY\_NOVELTY$, $JURY\_IMPORTANCE$, $JURY\_RELEVANCE$, $JURY\_SCALABILITY$.

Рассмотрим получившиеся факторы и их нагрузки (табл. 2.1).

Таблица 2.1. Факторы.

Variable	Factor Loadings (Unrotated) (KIVO Data_2.s Extraction: Principal components (Marked loadings are >,700000) Factor 1	Factor 2
JURY NOVELTY	-0,666980	0,607738
JURY IMPORTANCE	-0,794594	0,218185
JURY RELEVANCE	-0,807517	-0,239089
JURY SCALABILITY	-0,690036	-0,558882
Expl.Var	2,204475	0,786463
Prp.Totl	0,551119	0,196616

Видно, что достаточно интерпретировать первый фактор. Первый фактор теснее всего связан с $JURY\_IMPORTANCE$ — актуальность решаемых проблем и $JURY\_RELEVANCE$ — целесообразность используемых механизмов. Фактор новизны и удобства, а новизна и удобство противоположны: либо делается что-то новое, либо хорошо делается старое.

Далее проверялось, можно ли считать, что эксперты примерно одинаково оценивают каждый проект по этим четырем признакам, т.е если высокая оценка за актуальность, то высокие оценки и по другим критериям. Для первого фактора:

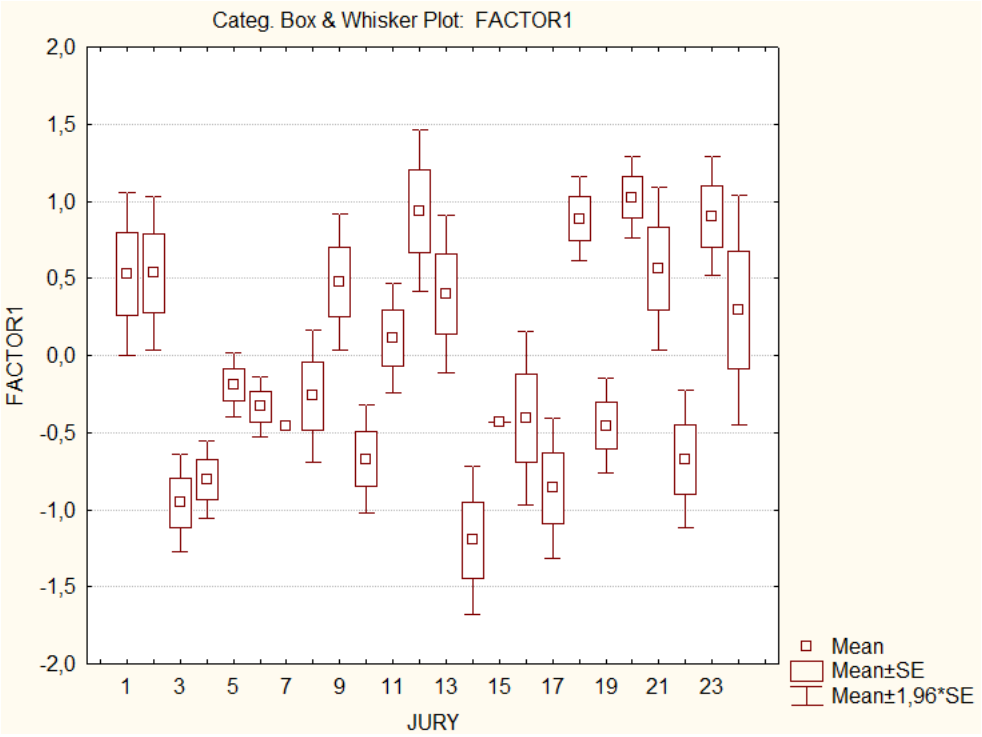

Рис. 2.1. Диаграмма размаха для $Factor1$ и $JURY$.

По диаграмме на рис. 2.1 видно, что одни эксперты занижают (те, у которых фактор наверху), а другие завышают. Чем больше первый фактор, тем больше эксперт занижает. В общем эксперты оценивают достаточно адекватно, т.е можно соединить 4 признака в одну оценку.

Второй фактор тоже немаловажен. На рис. 2.2 видно, что есть эксперты, который завышают $JURY\_NOVELTY$ и $JURY\_IMPORTANCE$, а занижают $JURY\_RELEVANCE$ и $JURY\_SCALABILITY$, т.е эксперт завышает оценку тому проекту, который удовлетворяет его предпочтениям.

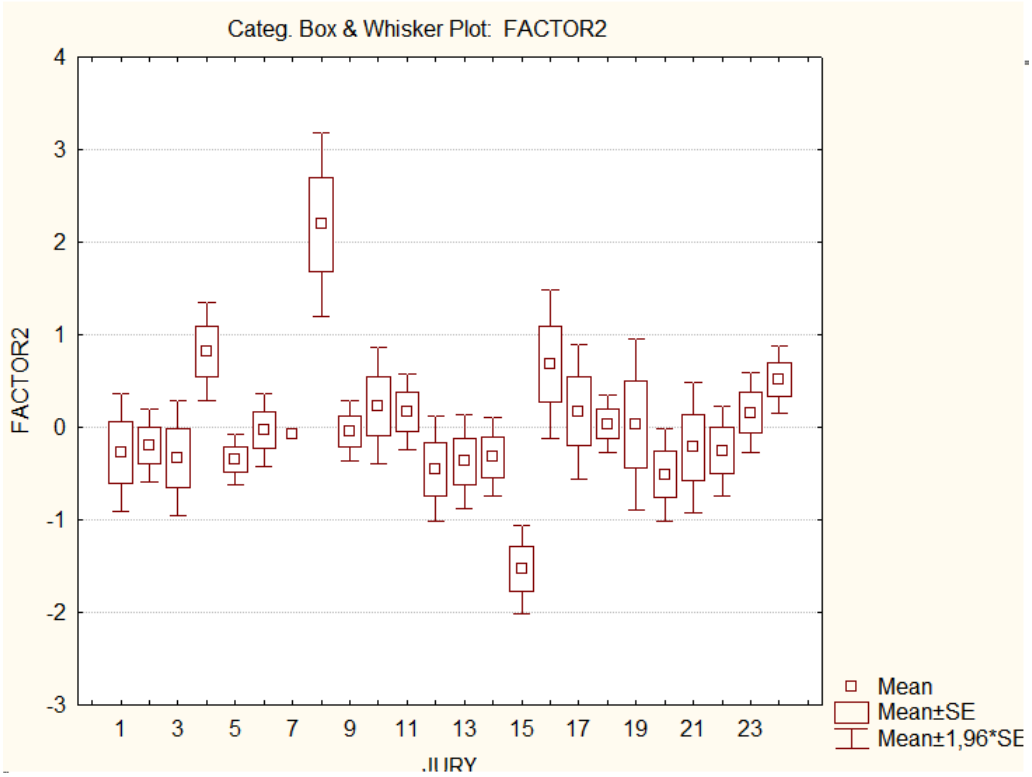

Рис. 2.2. Диаграмма размаха для $Factor2$ и $JURY$.

Благодаря этой диаграмме можно «вытащить» неадекватных экспертов. Видно, что это эксперт номер 8 и номер 15.

2.1.2. Качество оценивания выживаемости экспертами

Рассмотрен признак $JURY\_OVERALL$ — общая оценка эксперта (3 группы) и два получившихся фактора.

Произведено сравнение в трех группах:

0 — далее не рассматривать,

1 — рассматривать в общем порядке,

2 — обратить особое внимание,

т.е 0 — проект отвергается, 1 или 2 — принимается во внимание.

Был проведен однофакторный дисперсионный анализ.

В табл. 2.2 основные результаты анализа: суммы квадратов, степени свободы, значения

Таблица 2.2. Однофакторный дисперсионный анализ.

Analysis of Variance (Spreadsheet8.sta)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
FACTOR1	39,69814	1	39,69814	234,3019	273	0,858249	46,25482	0,000000

Analysis of Variance (Spreadsheet8.sta)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
FACTOR2	0,089254	1	0,089254	273,9107	273	1,003336	0,088957	0,765734

F-критерия, уровни значимости.

Для удобства исследования значимые эффекты (p<0.05) выделены красным цветом.

$Factor1$ получился значимым, т.е он влияет на разделение по группам.

Необходимо проверить, хорошо прогнозируют эксперты или нет. Были ли они правы в своих предсказаниях? Рассмотрены два качественных признака $JURY\_OVERALL$ — оценка эксперта и $SURVIVE$ — Продолжают ли работу над проектом (0 — нет, 1 — да).

С помощью статистики хи-квадрат проверена гипотеза о наличии взаимосвязи между двумя качественными признаками:

```
 Pearson's Chi-squared test

data:  sort
X-squared = 0.015, df = 1, p-value = 0.9025
```

Зависимости нет, можно сделать вывод, что эксперты не определяют выживет проект или нет.

Был рассмотрен признак выжил/не выжил проект $SURVIVE$ и $Factor1$, $Factor2$. Получено, что факторы не влияют на успешность проекта. Итог не прогнозируется экспертами (табл. 2.3).

Таблица 2.3. Дисперсионный анализ $SURVIVE$ и $Factor1$, $Factor2$.

Analysis of Variance (KIVO Data_2)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Factor1	0,134609	1	0,134609	273,8654	273	1,003170	0,134183	0,714417
Factor2	0,036950	1	0,036950	273,9630	273	1,003528	0,036820	0,847974

2.1.3. Изолированный анализ качественных признаков «До»

Изучается влияние блока информации «До» на итоговую характеристику $SURVIVE$ — Продолжают ли работу над проектом (0 — нет, 1 — да).

Были найдены зависимости с помощью критерия Хи-квадрат и упорядочены по убыванию влияния те признаки, которые имеют связь с $SURVIVE$ (чем меньше p, тем больше влияние)(табл. 2.4).

Таблица 2.4. Связь $SURVIVE$ с признаками «До».

AGE	$p - value <$ 2.2e – 16 — сколько лет работают над проектом.
$WEB.SITE$	$p - value =$ 1.706e – 05 — наличие веб-сайта.
$INDUSTRY$	$p - value =$ 0.001957 — есть ли в команде сотрудник организации.
$FAMILY$	$p - value =$ 0.006592 — входит ли семейное образование в сферу проекта.
$TEACHER$	$p-value =$ 0.0068 — есть ли в команде преподаватель.
$KINDERGARTEN$	$p - value =$ 0.02128 — входит ли дошкольное образование в сферу проекта.

Таким образом получены признаки, каждый из которых (в одиночку) оказывает влияние на итоговую характеристику.

2.1.4. Поиск наилучшего подмножества с помощью коэффициента неопределенности

Необходимо найти подмножество признаков, которое наибольшим образом связано с итоговой характеристикой $SURVIVE$.

(Наложено ограничение на кол-во элементов в подмножестве: не более трех признаков).

Назовем его номинативным представителем множества признаков «До» для упрощенной задачи поиска связи между одним признаком «После» и множеством «До».

Были посчитаны коэффициенты неопределенности и выделены связанные с $SURVIVE$ подмножества (табл. 2.5, табл. 2.6 и табл. 2.7).

Таблица 2.5. Подмножества, состоящие из одного признака.

AGE	J= 36.95242

Таблица 2.6. Подмножества, состоящие из двух признаков.

$OS + AGE$	J=28.93851
$LOCATION + AGE$	J=24.45135
$KINDERGARTEN + AGE$	J=29.56132
$PRIMARY.SCHOOL + AGE$	J=26.53827
$MIDDLE.SCHOOL + AGE$	J=25.26018
$HIGH.SCHOOL + AGE$	J=25.41795
$UNIVERSITY + AGE$	J=25.1947
$EXTRACURRICULAR + AGE$	J=25.12047
$PROFESSIONA + AGE$	J=25.47822
$FAMILY + AGE$	J=27.65472
$OTHER + AGE$	J=28.85889
$WEB.SITE + AGE$	J=25.82245
$SEX + AGE$	J=21.33207
$TEACHER + AGE$	J=26.05954
$ENREPRENEUR + AGE$	J=29.41551
$INDUSTRY + AGE$	J=26.36016
$AGE + TEAM.NUM$	J=26.82852

Таблица 2.7. Подмножества, состоящие из трех признаков.

$OS + KINDERGARTEN + AGE$	J=24.45471

$OS + PRIMARY.SCHOOL + AGE$	J=22.1477
$OS + FAMILY + AGE$	J=22.97665
$OS + OTHER + AGE$	J=23.79405
$OS + ENREPRENEUR + AGE$	J=24.13174
$OS + INDUSTRY + AGE$	J=22.09018
$OS + AGE + TEAM.NUM$	J=22.68196
$KINDERGARTEN + PRIMARY.SCHOOL + AGE$	J=22.69849
$KINDERGARTEN + FAMILY + AGE$	J=23.36680
$KINDERGARTEN + OTHER + AGE$	J=23.62842
$KINDERGARTEN + TEACHER + AGE$	J=22.15628
$KINDERGARTEN + ENREPRENEUR + AGE$	J=24.45947
$KINDERGARTEN + INDUSTRY + AGE$	J=22.15707
$KINDERGARTEN + AGE + TEAM.NUM$	J=22.77684
$PRIMARY.SCHOOL + ENREPRENEUR + AGE$	J=22.32812
$FAMILY + OTHER + AGE$	J=22.5581
$FAMILY + ENREPRENEUR + AGE$	J=23.20166
$OTHER + ENREPRENEUR + AGE$	J=23.8796
$OTHER + INDUSTRY + AGE$	J=22.02824
$OTHER + AGE + TEAM.NUM$	J=22.47287
$WEB.SITE + ENREPRENEUR + AGE$	J=22.16894
$ENREPRENEUR + INDUSTRY + AGE$	J=22.48607
$ENREPRENEUR + AGE + TEAM.NUM$	J=23.14652

Рассмотрим таблички (для наглядности), в которых указано, сколько раз встречаются признаки «До» в подмножествах и с какими коэффициентами неопределенности (табл. 2.8, табл. 2.9 и табл. 2.10).

Таблица 2.8. Подмножества, состоящие из одного признака.

	name	I<1%	1%<I<10%	I>10%
1	OS	1	0	0
2	LOCATION	1	0	0
3	KINDERGARTEN	1	0	0
4	PRIMARY.SCHOOL	1	0	0
5	MIDDLE.SCHOOL	1	0	0
6	HIGH.SCHOOL	1	0	0
7	UNIVERSITY	1	0	0
8	EXTRACURRICULAR	1	0	0
9	PROFESSIONAL	1	0	0
10	FAMILY	0	1	0
11	OTHER	1	0	0
12	WEB.SITE	0	1	0
13	SEX	1	0	0
14	TEACHER	0	1	0
15	ENREPRENEUR	1	0	0
16	INDUSTRY	0	1	0
17	AGE	0	0	1
18	TEAM.NUM	1	0	0

Таблица 2.9. Подмножества, состоящие из двух признаков.

	name	I<1%	1%<I<10%	I>10%
1	OS	4	12	1
2	LOCATION	12	4	1
3	KINDERGARTEN	10	6	1
4	PRIMARY.SCHOOL	14	2	1
5	MIDDLE.SCHOOL	13	3	1
6	HIGH.SCHOOL	14	2	1
7	UNIVERSITY	12	4	1
8	EXTRACURRICULAR	14	2	1
9	PROFESSIONAL	13	3	1
10	FAMILY	7	9	1
11	OTHER	12	4	1
12	WEB.SITE	0	16	1
13	SEX	12	4	1
14	TEACHER	8	8	1
15	ENREPRENEUR	11	5	1
16	INDUSTRY	2	14	1
17	AGE	0	0	17
18	TEAM.NUM	6	10	1

Таблица 2.10. Подмножества, состоящие из трех признаков.

	name	I<1%	1%<I<10%	I>10%
1	OS	3	117	16
2	LOCATION	39	81	16
3	KINDERGARTEN	32	88	16
4	PRIMARY.SCHOOL	58	62	16
5	MIDDLE.SCHOOL	60	60	16
6	HIGH.SCHOOL	55	65	16
7	UNIVERSITY	45	75	16
8	EXTRACURRICULAR	62	58	16
9	PROFESSIONAL	49	71	16
10	FAMILY	22	98	16
11	OTHER	45	75	16
12	WEB.SITE	0	120	16
13	SEX	35	85	16
14	TEACHER	24	96	16
15	ENREPRENEUR	48	72	16
16	INDUSTRY	10	110	16
17	AGE	0	0	136
18	TEAM.NUM	1	119	16

Можно сделать вывод, что AGE является «номинативным представителем». При его добавлении к другим признакам, получаем подмножества, влияющие на итоговую характеристику $SURVIVE$. Посмотрим на таблицу сопряженности 2.11 AGE и $SURVIVE$, чтобы узнать, какие проекты на самом деле выживают:

Таблица 2.11. Таблица сопряженности AGE и $SURVIVE$.

AGE\SURVIVE	не выжил	выжил
0	312	0
1	0	145
2	35	0
3	12	0
4	48	0

Получилось, что выживают только проекты «1», т.е над которыми работали от 1 до 2 лет.

Было проверено, насколько хорошо признаки в подмножествах зависимы с помощью

критерия Хи-квадрат:

В подмножествах, состоящих из двух признаков, нет зависимости между признаками, а в подмножествах, состоящих из трех признаков, она есть.

Получились ковариационные триады.

Ковариационные триады — это недавнее изобретение. Ими занимался Юрий Белоусов под руководством Алексеевой Н. П.. Про парные ковариации событий упоминается в книге [6].

Рассмотрим, что это такое на примере бинарных признаков.

Пусть имеются три бинарных признака X_1, X_2, X_3. Соответственно обозначим через A_1, A_2, A_3 события, связанные с «успехами» $X_i = 1$, $i = 1, 2, 3$. Будем рассматривать ситуацию, при которой имеют место парные отрицательные ковариации

$$\mathrm{P}(A_i A_j) - \mathrm{P}(A_i)\,\mathrm{P}(A_j) < 0,\ \ i, j = 1, 2, 3$$

и положительная тройная ковариация

$$\mathrm{P}(A_1 A_2 A_3) - \mathrm{P}(A_1)\,\mathrm{P}(A_2)\,\mathrm{P}(A_3) > 0.$$

Нетрудно убедиться в том, что в таком случае

$$\mathrm{P}(A_1 A_2 \mid A_3) = \mathrm{P}(A_1 A_2 A_3) > \mathrm{P}(A_1)\,\mathrm{P}(A_2)\,\mathrm{P}(A_3) > \mathrm{P}(A_1 A_2)\,\mathrm{P}(A_3)$$

и для любой комбинации

$$\mathrm{P}(A_i A_j \mid A_k) > \mathrm{P}(A_i A_j)$$

условная вероятность произведения двух событий оказывается больше безусловной, т.е вероятность при добавлении третьего условия стала намного больше и A_3 выступает катализатором (одно условие, увеличивающее вероятность).

2.2. Множества признаков «До» и «После»

2.2.1. Перебор подмножеств и поиск наиболее связанных с помощью коэффициента неопределенности

Теперь будет решаться более сложная задача анализа связи между двумя множествами признаков «До» и «После».

Рассмотрим признаки «До» (X_m, $m = 6$):

AGE — Сколько лет Вы уже работаете над проектом, где

1) $AGE1$ — (1 — «до 2 лет», 0 — иначе),

2) $AGE2$ — (1 — «до 2 до 5 лет», 0 — иначе),

3) $AGE3$ — (1 — «от 5 лет», 0 — иначе),

4) $WEB.SITE$ — Наличие веб-сайта,

5) $TEACHER$ — Есть ли в команде преподаватель,

6) $PROFESSIONAL$ — Входит ли профессиональное образование в сферу проекта.

Рассмотрим признаки «После» (Y_n, $n = 5$):

1) $EVENTS.PUBLICATION$ — Публикации в СМИ о проекте,

2) $EVENTS.PARTNERS$ — Привлечены новые партнеры,

3) $EVENTS.GRANT$ — Получен грант,

4) $EVENTS.INVESTMENT$ — Привлечены инвестиции,

5) $FOREIGN$ — Есть ли в команде ключевые участники, имеющие заграничных опыт.

Найдены подмножества по алгоритму быстро перечисления точек грассманиана X_k, Y_k «До» и «После» размерности $k = 1$ и $k = 2$.

$\dim(X_k) = 1$: количество найденных симптомов $2^m - 1 = 2^6 - 1 = 64 - 1 = 63$.

$\dim(Y_k) = 1$: количество найденных симптомов $2^n - 1 = 2^5 - 1 = 31$.

$\dim(X_k) = 2$: количество найденных синдромов 651.

$\dim(Y_k) = 2$: количество найденных синдромов 155.

Выделены наиболее связанные подмножества, в качестве метрики используется односторонний коэффициент неопределенности:

Для X_k и Y_k, $k = 1$: 36 штук.

Для X_k и Y_k, $k = 2$: 740 штук.

2.2.2. Частотный способ поиска номинативных представителей

Построены таблицы частот 2.12, 2.13, 2.14 и 2.15 (сколько раз какой признак встречается в подмножествах):

Таблица 2.12. Частота признаков X в подмножествах $X_k,\ k = 1$.

name	frequency
X(1)	18
X(2)	18
X(3)	18
X(4)	0
X(5)	0
X(6)	0

Таблица 2.13. Частота признаков X в подмножествах $X_k,\ k = 2$.

name	frequency
X(1)	556
X(2)	556
X(3)	556
X(4)	512
X(5)	348
X(6)	324

Таблица 2.14. Частота признаков Y в подмножествах $Y_k,\ k = 1$.

name	frequency
Y(1)	32
Y(2)	12
Y(3)	16
Y(4)	18
Y(5)	12

Таблица 2.15. Частота признаков Y в подмножествах $Y_k,\ k = 2$.

name	frequency
Y(1)	740
Y(2)	180
Y(3)	507
Y(4)	483
Y(5)	268

Симптомы, из которых состоят наиболее связанные подмножества $\dim(X_k) = 1$ и $\dim(Y_k) = 1$ (табл. 2.16, табл. 2.17):

Таблица 2.16. Частота симптомов X_1.

	name	frequency
1	X(1)	15
2	X(2)	3
3	X(1)+X(3)	3
4	X(2)+X(3)	15

Таблица 2.17. Частота симптомов Y_1.

	name	frequency
1	Y(1)	4
2	Y(2)	2
3	Y(1)+Y(2)	2
4	Y(1)+Y(3)	4
5	Y(1)+Y(2)+Y(3)	2
6	Y(1)+Y(4)	2
7	Y(1)+Y(2)+Y(4)	2
8	Y(1)+Y(2)+Y(3)+Y(4)	2
9	Y(1)+Y(3)+Y(4)	4
10	Y(1)+Y(5)	2
11	Y(1)+Y(3)+Y(5)	2
12	Y(1)+Y(3)+Y(4)+Y(5)	2
13	Y(1)+Y(2)+Y(4)+Y(5)	2
14	Y(1)+Y(4)+Y(5)	2
15	Y(4)+Y(5)	2

Синдромы, из которых состоят подмножества $\dim(X_k) = 2$ и $\dim(Y_k) = 2$ (табл. 2.19, табл. 2.18):

Таблица 2.18. Частота синдромов Y_2.

	name	frequency
1	Y(1) and Y(3)	61
2	Y(1) and Y(2)+Y(3)	52
3	Y(1) and Y(4)	61
4	Y(1) and Y(3)+Y(4)	61
5	Y(1) and Y(5)	60
6	Y(1) and Y(4)+Y(5)	60
7	Y(1)+Y(3) and Y(4)	61
8	Y(1) and Y(2)	24
9	Y(1) and Y(2)+Y(4)	16
10	Y(1) and Y(3)+Y(5)	20
11	Y(1) and Y(3)+Y(4)+Y(5)	28
12	Y(2) and Y(1)+Y(3)	28
13	Y(2) and Y(1)+Y(3)+Y(4)	28
14	Y(3) and Y(1)+Y(4)	48
15	Y(1)+Y(3) and Y(1)+Y(4)	8
16	Y(1)+Y(3) and Y(1)+Y(5)	8
17	Y(1)+Y(3) and Y(4)+Y(5)	16
18	Y(2)+Y(3) and Y(1)+Y(2)+Y(4)	8
19	Y(1)+Y(4) and Y(1)+Y(5)	8
20	Y(1)+Y(3)+Y(4) and Y(1)+Y(5)	12
21	Y(1)+Y(3)+Y(4) and Y(2)+Y(3)+Y(5)	8
22	Y(1)+Y(3)+Y(4) and Y(1)+Y(3)+Y(5)	32
23	Y(1)+Y(3) and Y(5)	4
24	Y(1)+Y(3) and Y(1)+Y(4)+Y(5)	8
25	Y(1)+Y(3)+Y(4) and Y(3)+Y(5)	4
26	Y(1) and Y(2)+Y(3)+Y(4)	12
27	Y(2) and Y(1)+Y(4)	4

Таблица 2.19. Частота синдромов X_2.

	name	frequency
1	X(1) and X(2)	7
2	X(1) and X(3)	7
3	X(1) and X(2)+X(3)	4
4	X(1) and X(4)	22
5	X(1) and X(2)+X(4)	22
6	X(1) and X(2)+X(3)+X(4)	22
7	X(1) and X(3)+X(4)	22
8	X(1) and X(5)	10
9	X(1) and X(2)+X(5)	7
10	X(1) and X(2)+X(3)+X(5)	10
11	X(1) and X(3)+X(5)	7
12	X(1) and X(3)+X(4)+X(5)	17
13	X(1) and X(2)+X(3)+X(4)+X(5)	18
14	X(1) and X(2)+X(4)+X(5)	17
15	X(1) and X(4)+X(5)	18
16	X(1) and X(6)	9
17	X(1) and X(2)+X(6)	8
18	X(1) and X(2)+X(3)+X(6)	9
19	X(1) and X(3)+X(6)	8
20	X(1) and X(3)+X(4)+X(6)	15
21	X(1) and X(2)+X(3)+X(4)+X(6)	14
22	X(1) and X(2)+X(4)+X(6)	15
23	X(1) and X(4)+X(6)	14
24	X(1) and X(4)+X(5)+X(6)	7
25	X(1) and X(2)+X(4)+X(5)+X(6)	13
26	X(1) and X(2)+X(3)+X(4)+X(5)+X(6)	7
27	X(1) and X(3)+X(4)+X(5)+X(6)	13
28	X(1) and X(3)+X(5)+X(6)	7
29	X(1) and X(2)+X(3)+X(5)+X(6)	8
30	X(1) and X(2)+X(5)+X(6)	7
31	X(1) and X(5)+X(6)	8
32	X(2) and X(3)	7
33	X(1)+X(2) and X(1)+X(3)	7
34	X(2)+X(3) and X(4)	22
35	X(2)+X(3) and X(1)+X(4)	22
36	X(2)+X(3) and X(1)+X(2)+X(4)	22
37	X(2)+X(3) and X(2)+X(4)	22
38	X(2)+X(3) and X(5)	10
39	X(2)+X(3) and X(1)+X(5)	10
40	X(2)+X(3) and X(1)+X(2)+X(5)	7
41	X(2)+X(3) and X(2)+X(5)	7
42	X(2)+X(3) and X(2)+X(4)+X(5)	17
43	X(2)+X(3) and X(1)+X(2)+X(4)+X(5)	17
44	X(2)+X(3) and X(1)+X(4)+X(5)	18
45	X(2)+X(3) and X(4)+X(5)	18
46	X(2)+X(3) and X(6)	9
47	X(2)+X(3) and X(1)+X(6)	9
48	X(2)+X(3) and X(1)+X(2)+X(6)	8
49	X(2)+X(3) and X(2)+X(6)	8
50	X(2)+X(3) and X(2)+X(4)+X(6)	15
51	X(2)+X(3) and X(1)+X(2)+X(4)+X(6)	15
52	X(2)+X(3) and X(1)+X(4)+X(6)	14
53	X(2)+X(3) and X(4)+X(6)	14
54	X(2)+X(3) and X(4)+X(5)+X(6)	7
55	X(2)+X(3) and X(1)+X(4)+X(5)+X(6)	7
56	X(2)+X(3) and X(1)+X(2)+X(4)+X(5)+X(6)	13
57	X(2)+X(3) and X(2)+X(4)+X(5)+X(6)	13
58	X(2)+X(3) and X(2)+X(5)+X(6)	7
59	X(2)+X(3) and X(1)+X(2)+X(5)+X(6)	7
60	X(2)+X(3) and X(1)+X(5)+X(6)	8
61	X(2)+X(3) and X(5)+X(6)	8

Поиск номинативных представителей:

По таблицам 2.13, 2.16 можно заметить, что признаки $AGE1, AGE2, AGE3$ образуют номинативный представитель множества «До».

По таблицам 2.15, 2.18 видно, что признак Y(1) $-EVENTS.PUBLICATION$ входит в номинативный представитель множества «После».

Посмотрим на информативность симптомов.

Были посчитаны и упорядочены по возрастанию энтропии симптомов в значимых подмножествах размерности 1(табл. 2.20, табл. 2.21).

Таблица 2.20. Энтропии симптомов в значимых подмножествах $X_k,\ k = 1$.

name	entropy
X(2)	0.5486287
X(1)+X(3)	0.5486287
X(1)	0.6542118
X(2)+X(3)	0.6542118

Таблица 2.21. Энтропии симптомов в значимых подмножествах $Y_k,\ k = 1$.

name	entropy
Y(4)+Y(5)	0.394731
Y(1)+Y(2)	0.5014441
Y(1)+Y(3)+Y(4)+Y(5)	0.5175828
Y(1)+Y(3)	0.5228691
Y(1)+Y(2)+Y(4)	0.5281098
Y(2)	0.5384569
Y(1)+Y(2)+Y(3)+Y(4)	0.5384569
Y(1)+Y(3)+Y(4)	0.5384569
Y(1)+Y(2)+Y(4)+Y(5)	0.5384569
Y(1)+Y(4)+Y(5)	0.5384569
Y(1)+Y(3)+Y(5)	0.5536501
Y(1)+Y(2)+Y(3)	0.5733185
Y(1)+Y(4)	0.5781336
Y(1)+Y(5)	0.5923414
Y(1)	0.6197248

Рассмотрим во всех найденных подмножествах энтропии симптомов/синдромов и найдем минимальные (табл. 2.22, табл. 2.23 и табл. 2.24). Симптомы с маленькой энтропией означают, что признаки, образующие их, совпадают.

Таблица 2.22. Минимальные энтропии симптомов $X_k,\ k = 1$.

name	entropy
X(1)+X(2)	0.249882
X(3)	0.249882

Таблица 2.23. Минимальные энтропии симптомов $Y_k,\ k = 1$.

name	entropy
Y(3)	0.2416273
Y(4)	0.1983736
Y(3)+Y(4)	0.2817615

Таблица 2.24. Минимальные энтропии синдромов $X_k,\ k = 2$.

name	entropy
X(1)+X(2) and X(3)	0.249882

У синдромов подмножеств $Y_k,\ k = 2$ энтропия не близка к нулю.

Для наглядности были построены графики, где по оси x отложена энтропия симптомов/синдромов, из которых состоят наиболее связанные подмножества, а по оси y — частота появления симптомов/синдромов в этих подмножествах (рис. 2.3, 2.4, 2.5, 2.6).

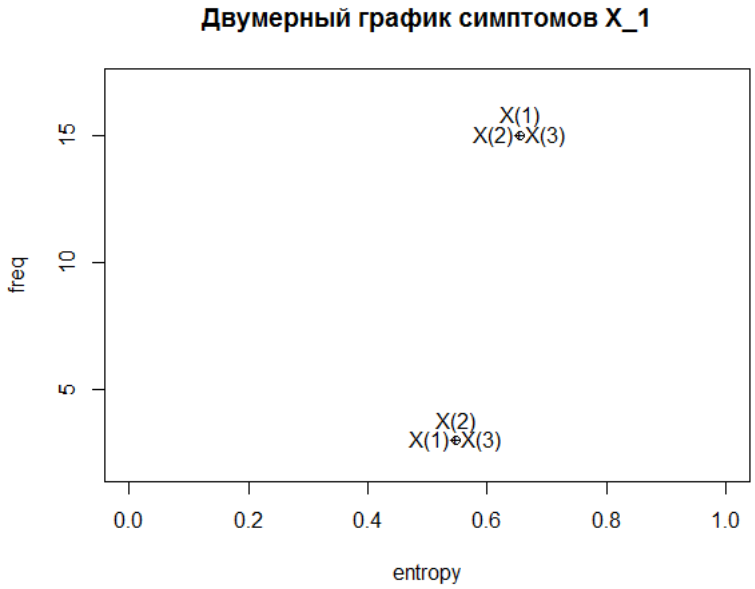

Рис. 2.3. Двумерный график симптомов X_1.

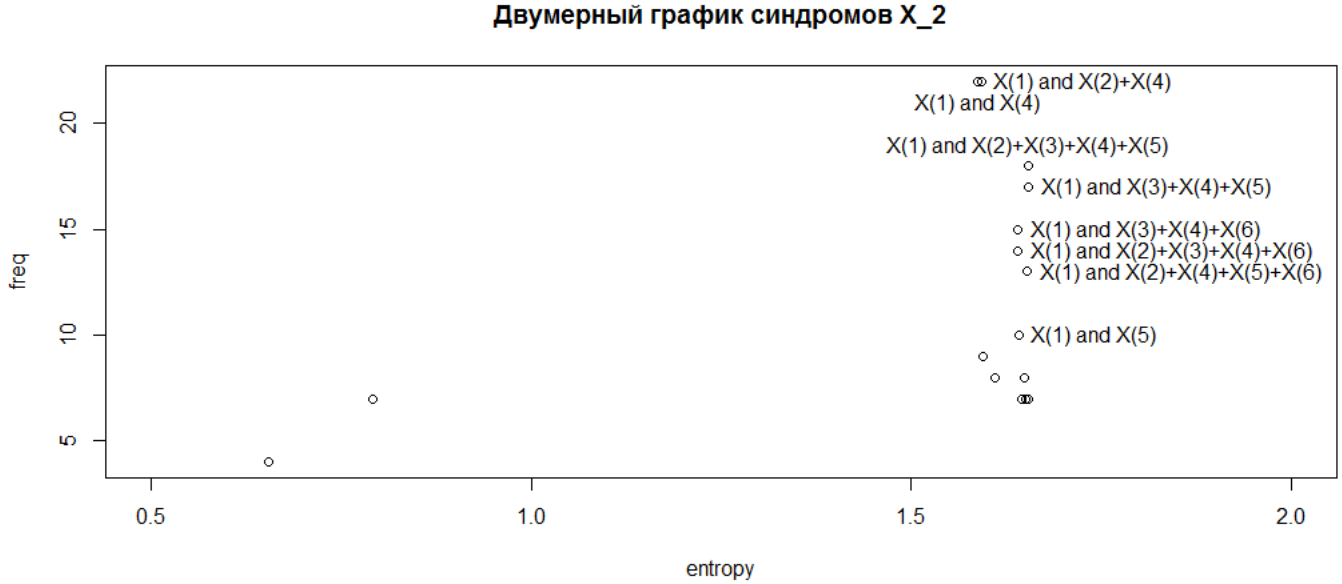

Рис. 2.4. Двумерный график синдромов X_2.

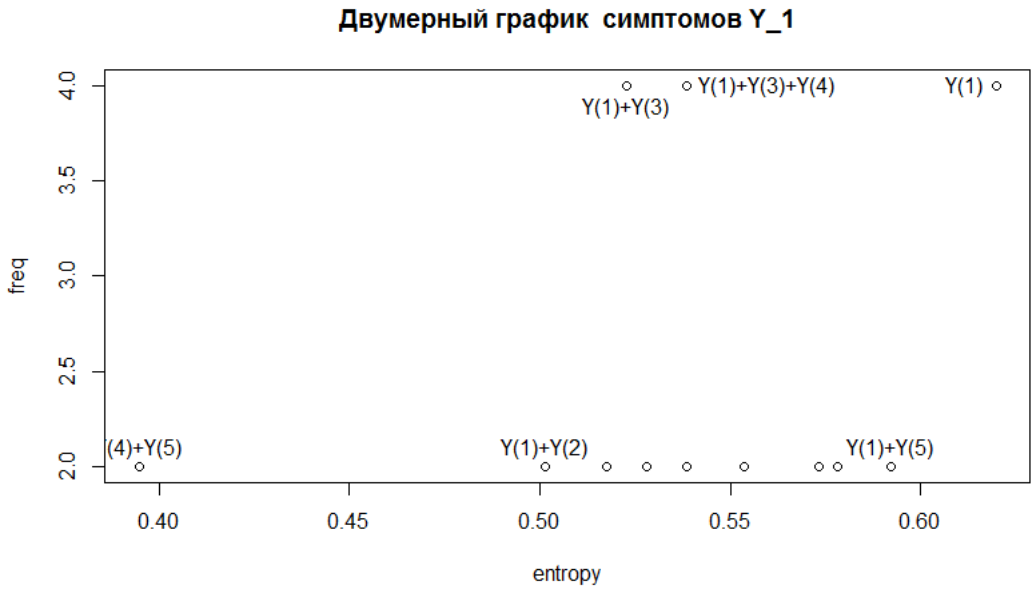

Рис. 2.5. Двумерный график симптомов Y_1.

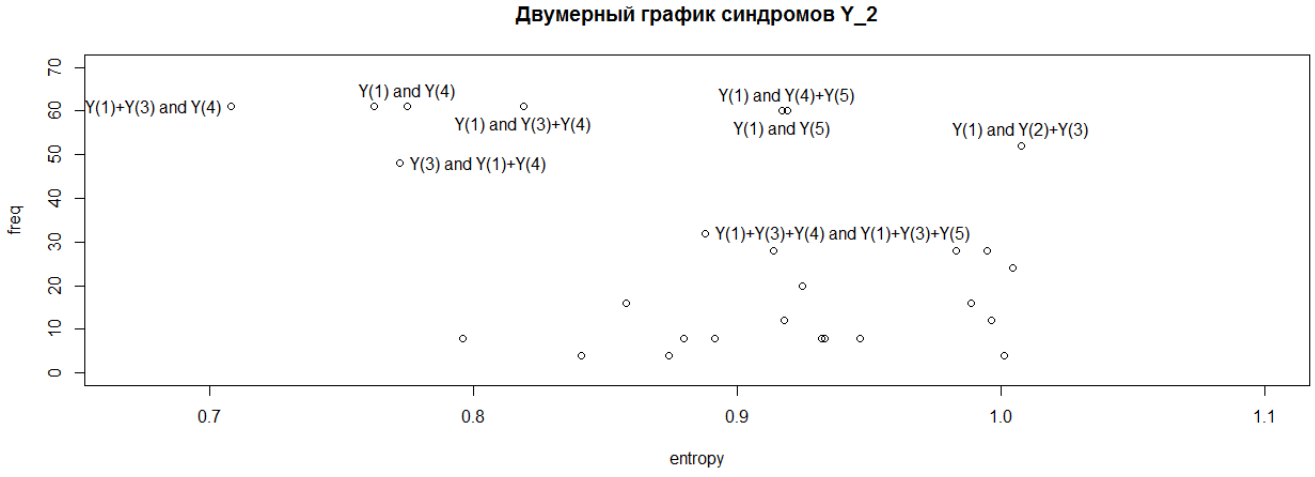

Рис. 2.6. Двумерный график синдромов Y_2.

Теперь посмотрим на двумерные графики, на которых изображены точками все симптомы в значимых подмножествах $X_k,\ Y_k, k = 2$ (рис. 2.7 и 2.8).

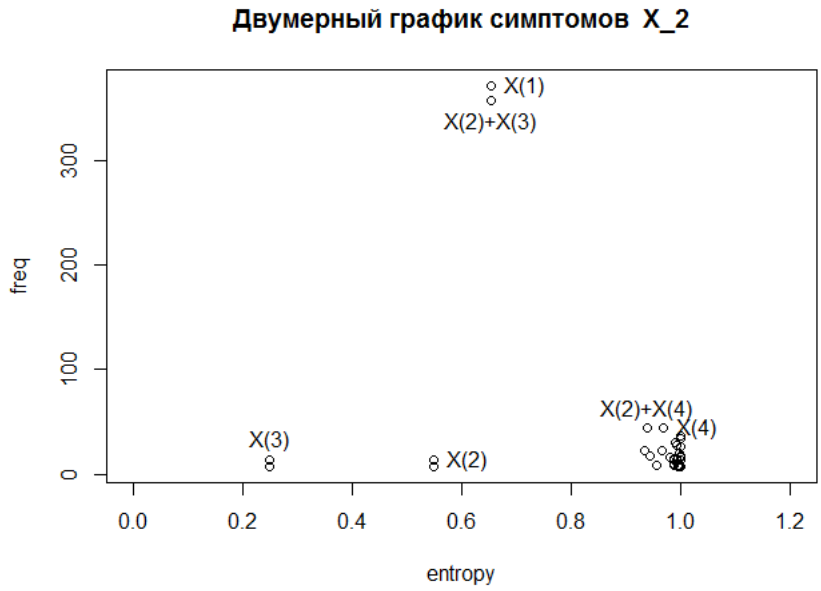

Рис. 2.7. Двумерный график симптомов, из которых состоят синдромы X_2.

Рис. 2.8. Двумерный график симптомов, из которых состоят синдромы Y_2.

2.2.3. Метод поиска номинативных представителей, основанный на удалении признаков

Рассмотрим связанные подмножества X_k Y_k, $k = 2$, получившиеся из исходных множеств X, Y. Будем удалять симптомы, которые содержатся в X_k, из этих подмножеств, оставляя неизменными Y_k, получим $\tilde{X}_k$. Проделаем аналогичные действия с симптомами, входящими в подмножества Y_k.

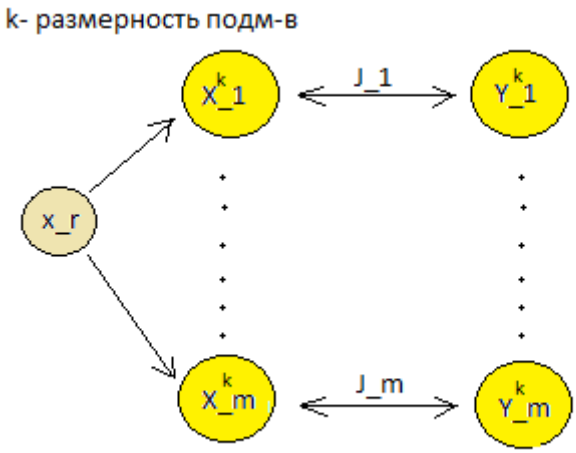

Рис. 2.9. До удаления симптома x_r.

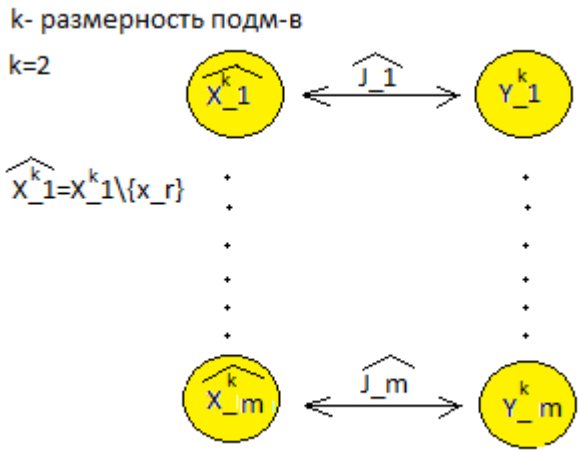

Рис. 2.10. После удаления симптома x_r.

Сравним коэффициенты неопределенности J и $\tilde{\text{J}}$ между подмножествами до удаления симптома X_k, Y_k и между $\tilde{X}_k, Y_k$. В ходе анализа были получены следующие таблицы (табл. 2.25, табл. 2.26 и табл. 2.27):

Расшифровка таблиц:

names — симптомы,

entr — энтропия симптомов,

freq_X — частоты встречаемости симптомов в значимых подмножествах,

without_X — сколько значимых подмножеств останется после удаления симптома,

means_J — среднее значение разности коэффициентов неопределенности между подмножествами до удаления симптома и после,

freq_Y — частоты признаков Y , в подмножествах Y_k, которые связаны с X_k, содержащими этот симптом.

Таблица 2.25. Таблица статистик для X_k, $k = 2$.

names	entr	freq_X	without_X	means_J	freq_Y(1)	freq_Y(2)	freq_Y(3)	freq_Y(4)	freq_Y(5)
X(1)	0.654211817176747	372	10	25.2728061	556	135	381	363	201
X(2)	0.54862867553	14	7	12.5106468	556	135	381	363	201
X(3)	0.249882292833185	14	13	3.5454152	556	135	381	363	201
X(2)+X(3)	0.654211817176747	358	4	25.6718529	446	110	308	293	163
X(4)	0.967905948648272	44	44	1.7351111	512	152	364	336	200
X(2)+X(4)	0.938516792157484	44	44	2.1851151	201	60	144	132	78
X(2)+X(3)+X(4)	0.933744057620171	22	22	1.7351111	61	18	43	40	24
X(3)+X(4)	0.964471424314364	22	22	2.1851151	128	38	91	84	50
X(5)	0.997261585661256	20	20	0.9009127	348	100	224	228	96
X(2)+X(5)	0.992021443918131	14	14	0.5305879	42	3	24	27	12
X(2)+X(3)+X(5)	0.987692508895803	10	10	0.9009127	18	3	12	12	5
X(3)+X(5)	0.994538681650011	7	7	0.5305879	32	4	20	21	9
X(3)+X(4)+X(5)	0.99965907166914	17	17	1.2901428	55	21	36	36	15
X(2)+X(3)+X(4)+X(5)	0.99965907166914	18	18	1.2782781	25	8	17	17	8
X(2)+X(4)+X(5)	0.999763249914076	34	34	1.2901428	90	39	57	57	21
X(4)+X(5)	0.999763249914076	36	36	1.2782781	220	84	144	144	60
X(6)	0.943103436392458	18	18	0.8293738	324	60	216	212	108
X(2)+X(6)	0.978979616125383	16	16	0.6501938	24	3	15	15	6
X(2)+X(3)+X(6)	0.986282531380956	9	9	0.8293738	9	1	6	6	4
X(3)+X(6)	0.955758991215001	8	8	0.6501938	17	2	11	11	6
X(3)+X(4)+X(6)	0.987692508895803	15	15	1.3677589	29	6	22	19	13
X(2)+X(3)+X(4)+X(6)	0.986282531380956	14	14	1.3680365	14	3	10	8	6
X(2)+X(4)+X(6)	0.989663018374806	30	30	1.3677589	45	9	36	33	21
X(4)+X(6)	0.99088059340086	28	28	1.3680365	116	24	88	76	52
X(4)+X(5)+X(6)	0.999962121726866	14	14	0.8512777	80	24	48	52	16
X(2)+X(4)+X(5)+X(6)	0.99965907166914	26	26	1.0649712	39	18	24	24	6
X(2)+X(3)+X(4)+X(5)+X(6)	0.999990530493873	7	7	0.8512777	7	0	4	5	2

По таблице 2.25 и 2.26 видно, что нельзя удалить симптомы X(1), X(2)+X(3) и X(1)+X(3), иначе теряем значимые подмножества. При удалении X(1) остаются связи, за счет X(2), при удалении X(2)+X(3) остаются связи, за счет X(1).

Нельзя удалять признак Y(1), потому что он сильнее всего связан с симптомами из X_k. По этой же таблице можно обнаружить, что признаки Y(3) и Y(4) тоже достаточно часто встречаются в значимых подмножествах с симптомами из X_k.

Рассмотрев таблицу 2.27, получаем, что нельзя удалить симптомы Y(1), Y(1)+Y(3), Y(1)+Y(4), иначе теряем значимые подмножества.

Таблица 2.26. Таблица статистик для X_k, $k = 2$.

X(3)+X(4)+X(5)+X(6)	0.999763249914076	13	13	1.0649712	20	6	12	13	4
X(3)+X(5)+X(6)	0.996578810589144	7	7	0.4843961	15	1	9	10	4
X(2)+X(3)+X(5)+X(6)	0.99498482818597	8	8	0.6820769	8	1	5	5	2
X(2)+X(5)+X(6)	0.998143184120301	14	14	0.4843961	21	0	12	15	6
X(5)+X(6)	0.999052844165967	16	16	0.6820769	140	28	84	92	32
X(1)+X(2)	0.249882292833185	7	6	6.1033772	96	23	65	62	34
X(1)+X(3)	0.54862867553	7	0	24.0338405	7	1	4	4	2
X(1)+X(4)	0.933744057620171	22	22	1.7351111	61	18	43	40	24
X(1)+X(2)+X(4)	0.964471424314364	22	22	2.1851151	67	20	48	44	26
X(1)+X(5)	0.987692508895803	10	10	0.9009127	18	3	12	12	5
X(1)+X(2)+X(5)	0.994538681650011	7	7	0.5305879	14	1	8	9	4
X(1)+X(2)+X(4)+X(5)	0.99965907166914	17	17	1.2901428	30	13	19	19	7
X(1)+X(4)+X(5)	0.99965907166914	18	18	1.2782781	25	8	17	17	8
X(1)+X(6)	0.986282531380956	9	9	0.8293738	9	1	6	6	4
X(1)+X(2)+X(6)	0.955758991215001	8	8	0.6501938	8	1	5	5	2
X(1)+X(2)+X(4)+X(6)	0.987692508895803	15	15	1.3677589	15	3	12	11	7
X(1)+X(4)+X(6)	0.986282531380956	14	14	1.3680365	14	3	10	8	6
X(1)+X(4)+X(5)+X(6)	0.999990530493873	7	7	0.8512777	7	0	4	5	2
X(1)+X(2)+X(4)+X(5)+X(6)	0.999763249914076	13	13	1.0649712	13	6	8	8	2
X(1)+X(2)+X(5)+X(6)	0.996578810589144	7	7	0.4843961	7	0	4	5	2
X(1)+X(5)+X(6)	0.99498482818597	8	8	0.6820769	8	1	5	5	2

Таблица 2.27. Таблица статистик для Y_k, $k = 2$.

names	entr	freq_Y	without_Y	means_J	freq_X(1)	freq_X(2)	freq_X(3)	freq_X(4)	freq_X(5)	freq_X(6)
Y(1)	0.619724817256218	455	128	8.666551	556	556	556	512	348	324
Y(3)	0.241627306744078	109	109	-4.021573	381	381	381	364	224	216
Y(2)+Y(3)	0.473572924917452	60	60	-6.545591	60	60	60	56	44	28
Y(4)	0.198373665490835	122	122	-4.670040	363	363	363	336	228	212
Y(3)+Y(4)	0.281761577885065	61	61	-6.037417	139	139	139	140	84	72
Y(5)	0.347816913571063	64	64	-7.074804	201	201	201	200	96	108
Y(4)+Y(5)	0.394730971081214	76	76	-6.509223	84	84	84	80	40	56
Y(1)+Y(3)	0.522869066595288	133	72	5.647169	163	163	163	176	76	80
Y(2)	0.538456920463084	84	84	-6.436634	135	135	135	152	100	60
Y(2)+Y(4)	0.490445197273352	16	16	-7.648924	18	18	18	24	16	4
Y(3)+Y(5)	0.388238977512362	24	24	-7.441684	48	48	48	56	24	16
Y(3)+Y(4)+Y(5)	0.407518566568877	28	28	-7.530222	21	21	21	24	8	12
Y(1)+Y(3)+Y(4)	0.538456920463084	84	84	2.490866	63	63	63	72	32	24
Y(1)+Y(4)	0.578133611929811	68	20	7.899506	57	57	57	60	32	32
Y(1)+Y(5)	0.592341370373232	28	28	-4.877418	21	21	21	28	0	4
Y(1)+Y(2)+Y(4)	0.528109803462829	8	8	5.034355	6	6	6	8	4	0
Y(2)+Y(3)+Y(5)	0.484871721053115	8	8	-7.118254	6	6	6	8	4	0
Y(1)+Y(3)+Y(5)	0.553650147397331	32	32	-5.833298	24	24	24	24	12	12
Y(1)+Y(4)+Y(5)	0.538456920463084	8	8	-6.291080	6	6	6	8	0	4
Y(2)+Y(3)+Y(4)	0.490445197273352	12	12	-7.322351	9	9	9	12	12	4

Рассмотри набор, состоящий из коэффициентов неопределенности J до удаления симптома из X_k, и набор, состоящий из $\tilde{\text{J}}$ после удаления симптома. Необходимо узнать, как изменился набор коэффициентов неопределенности: уменьшился, увеличился.

Теоретически получаем (утверждение 1), если удаляемый симптом не зависит от других симптомов в подмножестве, то разность между коэффициентами неопределенности должна быть больше нуля, а если удаляемый симптом еще не зависит от связанного с ним подмножества Y_K, то разность между коэффициентами неопределенности должна быть равна нулю.

Проверим на практике:
Для каждого симптома x_r, который проверяется на номинативного представителя, есть набор разностей коэффициентов неопределенности $(\text{J} - \tilde{\text{J}})_i$ $i = 1, \ldots, p$, где p — число значимых подмножеств, которые содержат x_r и из которых можно удалить этот симптом. Для оценки изменения $\tilde{\text{J}}$ используется критерий знаков. Он дает возможность установить, на сколько однонаправленно изменяются значения коэффициентов неопределенности при повторном измерении после удаления.

Проверяется гипотеза H_0 : вероятность успеха и неудачи одинакова,
Альтернатива H_1 : вероятность успеха больше, чем 0.5.

Критерий реализован на языке R: binom.test(x,n,$alternative$), где x — число успехов, n — число испытаний, $alternative$ — альтернативная гипотеза. Успехом будем считать уменьшение $\tilde{\text{J}}$, т.е в качестве x берем количество положительных значений $\text{J} - \tilde{\text{J}}$. Результаты представлены в табл. 2.28.

Для всех наборов $(\text{J} - \tilde{\text{J}})_i$, где $i = 1, \ldots, p$, $p - value < 0.05$, поэтому отвергаем гипотезу H_0. Можно сделать вывод, что при удалении любого симптома из X_k , $\tilde{\text{J}}$ становится меньше коэффициента неопределенности J до удаления.

Таблица 2.28. Критерий знаков для симптомов $X_k,\ k = 2$.

	names	p.value
1	X(1)	8.31632781251602e-112
2	X(2)	6.103515625e-05
3	X(3)	6.103515625e-05
4	X(2)+X(3)	2.72509429760529e-107
5	X(4)	5.68434188608079e-14
6	X(2)+X(4)	5.68434188608079e-14
7	X(2)+X(3)+X(4)	2.38418579101562e-07
8	X(3)+X(4)	2.38418579101562e-07
9	X(5)	9.53674316406249e-07
10	X(2)+X(5)	6.103515625e-05
11	X(2)+X(3)+X(5)	0.0009765625
12	X(3)+X(5)	0.0078125
13	X(3)+X(4)+X(5)	7.62939453125e-06
14	X(2)+X(3)+X(4)+X(5)	3.814697265625e-06
15	X(2)+X(4)+X(5)	5.82076609134675e-11
16	X(4)+X(5)	1.45519152283668e-11
17	X(6)	3.814697265625e-06
18	X(2)+X(6)	0.0020904541015625
19	X(2)+X(3)+X(6)	0.001953125
20	X(3)+X(6)	0.03515625
21	X(3)+X(4)+X(6)	3.0517578125e-05
22	X(2)+X(3)+X(4)+X(6)	6.103515625e-05
23	X(2)+X(4)+X(6)	9.31322574615478e-10
24	X(4)+X(6)	3.72529029846191e-09
25	X(4)+X(5)+X(6)	6.103515625e-05
26	X(2)+X(4)+X(5)+X(6)	1.49011611938477e-08
27	X(2)+X(3)+X(4)+X(5)+X(6)	0.0078125
28	X(3)+X(4)+X(5)+X(6)	0.0001220703125
29	X(3)+X(5)+X(6)	0.0078125
30	X(2)+X(3)+X(5)+X(6)	0.00390625
31	X(2)+X(5)+X(6)	6.103515625e-05
32	X(5)+X(6)	1.52587890625e-05
33	X(1)+X(2)	0.0078125
34	X(1)+X(3)	0.0078125
35	X(1)+X(4)	2.38418579101562e-07
36	X(1)+X(2)+X(4)	2.38418579101562e-07
37	X(1)+X(5)	0.0009765625
38	X(1)+X(2)+X(5)	0.0078125
39	X(1)+X(2)+X(4)+X(5)	7.62939453125e-06
40	X(1)+X(4)+X(5)	3.814697265625e-06
41	X(1)+X(6)	0.001953125
42	X(1)+X(2)+X(6)	0.03515625
43	X(1)+X(2)+X(4)+X(6)	3.0517578125e-05
44	X(1)+X(4)+X(6)	6.103515625e-05
45	X(1)+X(4)+X(5)+X(6)	0.0078125
46	X(1)+X(2)+X(4)+X(5)+X(6)	0.0001220703125
47	X(1)+X(2)+X(5)+X(6)	0.0078125
48	X(1)+X(5)+X(6)	0.00390625

Воспользуемся критерием Вилкоксона для зависимых выборок [5], чтобы ответить на вопрос, значимы изменения $\tilde{\text{J}}$ или нет.

Проверяется гипотеза H_0 : выборки однородны,
Альтернатива H_1 : есть статистически значимое различие между выборками.

Критерий реализован на языке R: wilcox.test(x,y,$paired$), где x — J, y — $\tilde{\text{J}}$, $paired$ =TRUE — парный тест.

Результаты представлены в табл. 2.29.

Для всех наборов $(\text{J} - \tilde{\text{J}})_i$, где $i = 1, \ldots, p$, $p - value < 0.05$, поэтому отвергаем гипотезу H_0. Можно сделать вывод, что набор $\tilde{\text{J}}$ статистически значимо уменьшается по сравнению с исходным J, при удалении любого симптома из X_k .

Таблица 2.29. Тест Вилкоксона для симптомов $X_k,\ k = 2$.

	names	p.value
1	X(1)	3.19153589405577e-62
2	X(2)	0.0001220703125
3	X(3)	0.0001220703125
4	X(2)+X(3)	9.03483056503163e-60
5	X(4)	7.85461461025763e-09
6	X(2)+X(4)	7.86837113752951e-09
7	X(2)+X(3)+X(4)	4.76837158203125e-07
8	X(3)+X(4)	4.76837158203125e-07
9	X(5)	9.50217438815064e-05
10	X(2)+X(5)	0.00108773626806972
11	X(2)+X(3)+X(5)	0.001953125
12	X(3)+X(5)	0.015625
13	X(3)+X(4)+X(5)	1.52587890625e-05
14	X(2)+X(3)+X(4)+X(5)	7.62939453125e-06
15	X(2)+X(4)+X(5)	3.80240606486657e-07
16	X(4)+X(5)	1.74482181544845e-07
17	X(6)	0.000212246164657084
18	X(2)+X(6)	0.000850625603395708
19	X(2)+X(3)+X(6)	0.00390625
20	X(3)+X(6)	0.015625
21	X(3)+X(4)+X(6)	6.103515625e-05
22	X(2)+X(3)+X(4)+X(6)	0.0001220703125
23	X(2)+X(4)+X(6)	1.81624052193589e-06
24	X(4)+X(6)	3.98314908860449e-06
25	X(4)+X(5)+X(6)	0.00108464321894851
26	X(2)+X(4)+X(5)+X(6)	8.76704500766149e-06
27	X(2)+X(3)+X(4)+X(5)+X(6)	0.015625
28	X(3)+X(4)+X(5)+X(6)	0.000244140625
29	X(3)+X(5)+X(6)	0.015625
30	X(2)+X(3)+X(5)+X(6)	0.0078125
31	X(2)+X(5)+X(6)	0.00108155602458144
32	X(5)+X(6)	0.000478203901668711
33	X(1)+X(2)	0.015625
34	X(1)+X(3)	0.015625
35	X(1)+X(4)	4.76837158203125e-07
36	X(1)+X(2)+X(4)	4.76837158203125e-07
37	X(1)+X(5)	0.001953125
38	X(1)+X(2)+X(5)	0.015625
39	X(1)+X(2)+X(4)+X(5)	1.52587890625e-05
40	X(1)+X(4)+X(5)	7.62939453125e-06
41	X(1)+X(6)	0.00390625
42	X(1)+X(2)+X(6)	0.015625
43	X(1)+X(2)+X(4)+X(6)	6.103515625e-05
44	X(1)+X(4)+X(6)	0.0001220703125
45	X(1)+X(4)+X(5)+X(6)	0.015625
46	X(1)+X(2)+X(4)+X(5)+X(6)	0.000244140625
47	X(1)+X(2)+X(5)+X(6)	0.015625
48	X(1)+X(5)+X(6)	0.0078125

Проделаем аналогичные действия для симптомов Y_k: будем удалять каждый симптом y_r из Y_k, при фиксированном X_k, и проверять их на номинативных представителей. В этом случае доказанное утверждение 1 не работает, поэтому разность J и $\tilde{\text{J}}$ может быть любого знака.

Для каждого симптома y_r, который проверяется на номинативного представителя, есть набор разностей коэффициентов неопределенности $(\text{J} - \tilde{\text{J}})_i\ i = 1, \ldots, p$, где p — число значимых подмножеств, которые содержат y_r и из которых можно удалить этот симптом. Для оценки изменения $\tilde{\text{J}}$ используется критерий знаков.

Получено (табл. 2.30), что только для наборов $(\text{J} - \tilde{\text{J}})_i$, соответствующих симптомам Y(1), Y(1)+Y(3), Y(1)+Y(3)+Y(4), Y(1)+Y(4) и Y(1)+Y(2)+Y(4), $p - value < 0.05$ и отвергается гипотеза H_0. Можно сделать вывод, что при удалении любого из этих симптомов, $\tilde{\text{J}}$ становится меньше коэффициента неопределенности J до удаления.

Таблица 2.30. Критерий знаков для симптомов $Y_k,\ k = 2$.

name	p.value
Y(1)	1.07486017721073e-137
Y(1)+Y(3)	3.04374439692552e-17
Y(1)+Y(3)+Y(4)	5.16987882845644e-26
Y(1)+Y(4)	3.1047478687811e-08
Y(1)+Y(2)+Y(4)	0.00390625

Воспользуемся критерием Вилкоксона для зависимых выборок, чтобы ответить на вопрос, значимы изменения $\tilde{\text{J}}$ или нет. По таблице 2.31 видно, что для всех наборов $(\text{J} - \tilde{\text{J}})_i$, где $i = 1, \ldots, p$, $p - value < 0.05$, поэтому отвергаем гипотезу H_0.

Можно сделать вывод, что набор $\tilde{\text{J}}$ статистически значимо уменьшается по сравнению с исходным J, при удалении любого из симптомов Y(1), Y(1)+Y(3), Y(1)+Y(3)+Y(4), Y(1)+Y(4) и Y(1)+Y(2)+Y(4). Набор $\tilde{\text{J}}$ статистически значимо увеличивается по сравнению с исходным J, при удалении оставшихся симптомов: Y(3), Y(2)+Y(3), Y(4), Y(3)+Y(4), Y(5), Y(4)+Y(5), Y(2), Y(2)+Y(4), Y(3)+Y(5), Y(3)+Y(4)+Y(5), Y(1)+Y(5), Y(2)+Y(3)+Y(5), Y(1)+Y(3)+Y(5), Y(1)+Y(4)+Y(5), Y(2)+Y(3)+Y(4).

Таблица 2.31. Тест Вилкоксона для симптомов $Y_k,\ k = 2$.

	names	p.value
1	Y(1)	2.83150145814267e-76
2	Y(3)	1.29021777967993e-19
3	Y(2)+Y(3)	1.65437763763663e-11
4	Y(4)	9.31363693418388e-22
5	Y(3)+Y(4)	1.13142060130322e-11
6	Y(5)	3.581850146986e-12
7	Y(4)+Y(5)	3.65144906764186e-14
8	Y(1)+Y(3)	1.19609409650278e-18
9	Y(2)	1.73273377391311e-15
10	Y(2)+Y(4)	0.000469883076610589
11	Y(3)+Y(5)	1.89710960316386e-05
12	Y(3)+Y(4)+Y(5)	3.89509998303757e-06
13	Y(1)+Y(3)+Y(4)	1.73328810256325e-15
14	Y(1)+Y(4)	3.02058071857308e-10
15	Y(1)+Y(5)	3.92049583298986e-06
16	Y(1)+Y(2)+Y(4)	0.0133283287808176
17	Y(2)+Y(3)+Y(5)	0.0126438306738142
18	Y(1)+Y(3)+Y(5)	8.28332335650982e-07
19	Y(1)+Y(4)+Y(5)	0.0140290608983004
20	Y(2)+Y(3)+Y(4)	0.00238339823476365

Рассмотрим рис. 2.11 и рис. 2.12, на которых изображены ящики с усами для каждого удаляемого симптома из X_k (затем из Y_k), чтобы сравнить средние значения разностей $(\mathrm{J} - \tilde{\mathrm{J}})$ для этих симптомов. В номинативный представитель войдут симптомы с самыми высокими средними, а симптомы с небольшими средними значениями можно удалить из рассмотрения.

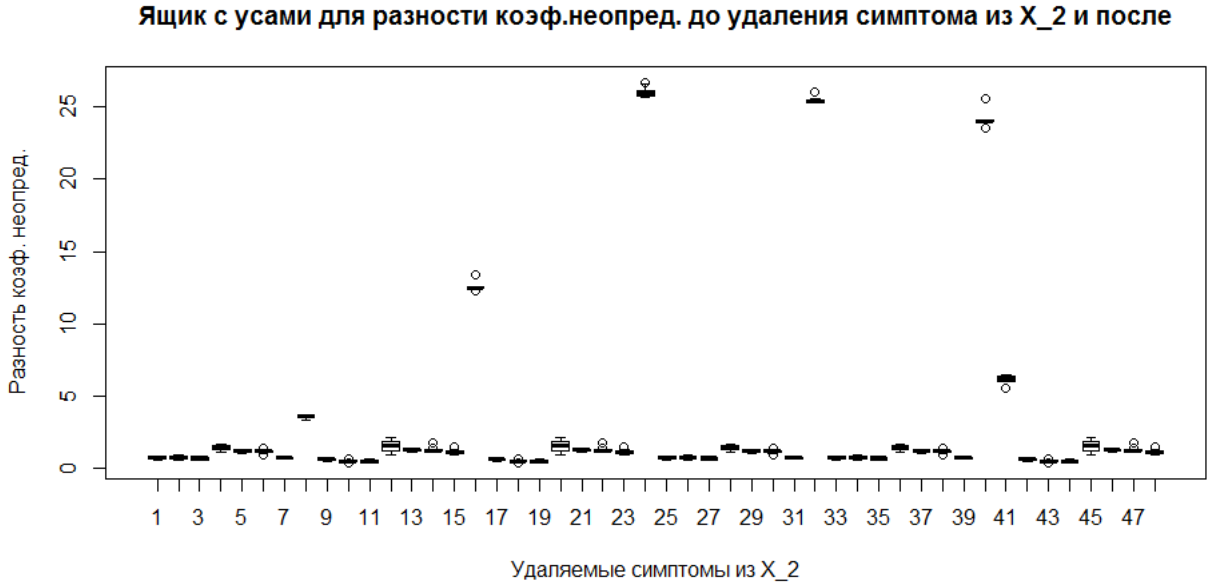

Рис. 2.11. Диаграмма размаха для разности коэффициентов неопределенности до удаления симптома из X_2 и после.

Симптомы X_k с самым большим средним значением разностей коэффициентов неопределенности до удаления симптома и после удаления:

№24 — X(2)+X(3),

№32 — X(1),

№40 — X(1)+X(3),

№16 — X(2),

№41 — X(1)+X(2).

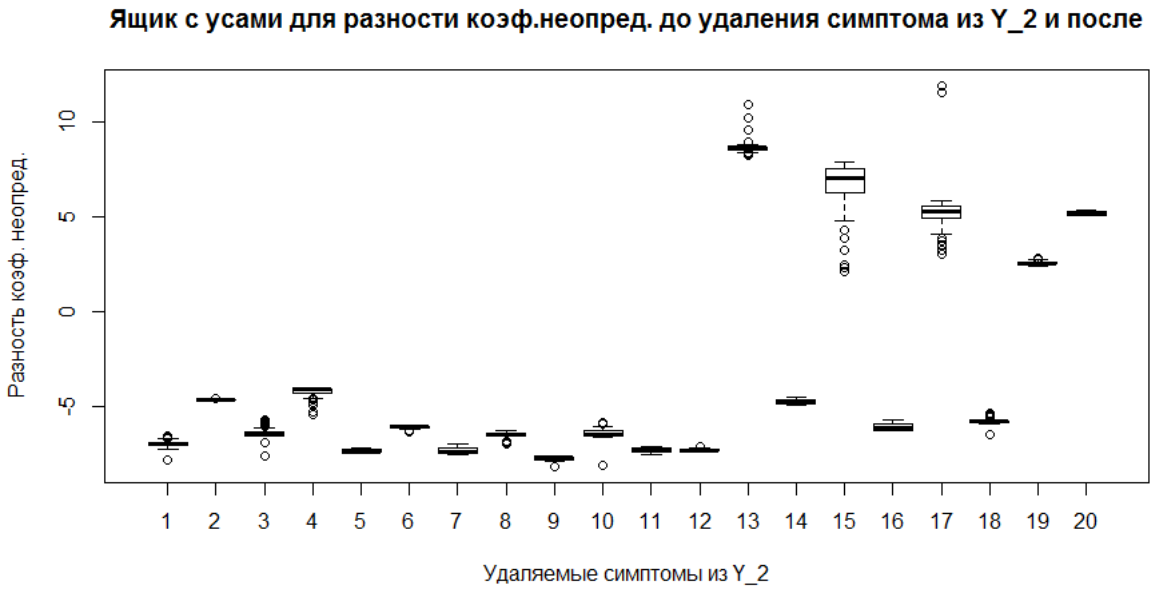

Рис. 2.12. Диаграмма размаха для разности коэффициентов неопределенности до удаления симптома из Y_2 и после.

Симптомы Y_k с самым большим средним значением разностей коэффициентов неопределенности до удаления симптома и после удаления:

№13 — Y(1),

№15 — Y(1)+Y(4),

№17 — Y(1)+Y(3),

№20 — Y(1)+Y(2)+Y(4),

№19 — Y(1)+Y(3)+Y(4).

2.2.4. Факторный анализ для поиска номинативных представителей

Проведем факторный анализ, используя таблицы 2.25, 2.26 и 2.27, чтобы найти устойчивые симптомы, в смысле уменьшения количества значимых связей и снижения уровней зависимости при их исключении из совокупностей.

Таблица 2.32. Матрица факторных нагрузок для симптомов X_k, $k = 2$.

	Comp.1	Comp.2	Comp.3	Comp.4
entr	0.2775282	-0.3060404	0.90859061	0.05638023
means_J	-0.2842746	0.5767919	0.26621416	-0.06872274
freq_Y.1.	-0.3935143	-0.2058274	0.04145998	0.12301059
freq_Y.2.	-0.3837639	-0.2439060	0.08931998	-0.84948918
freq_Y.3.	-0.3939258	-0.2044097	0.04130636	0.18071251
freq_Y.4.	-0.3933972	-0.2066318	0.04407521	0.13349216
freq_Y.5.	-0.3938885	-0.1897938	0.02627388	0.45063192
freq.X_without.X	-0.2737715	0.5927003	0.29926073	0.04229301

По таблице 2.32 видно, что нужно строить график по 2 и 3 компонентам, т.к они наиболее информативные (means_J — среднее значение разности коэффициентов неопределенности между подмножествами до удаления симптома и после, freq.X_without.X —разность между числом значимых подмножеств до удаления симптома и после , entr — энтропия симптомов). Для того, чтобы симптом входил в номинативный представитель нужно, чтобы means_J, freq.X_without.X и entr были наибольшими, поэтому берем симптомы, попадающие в верхний правый угол получившегося графика на рис. 2.13. Образуют устойчивые решения: X(1) и X(2)+X(3).

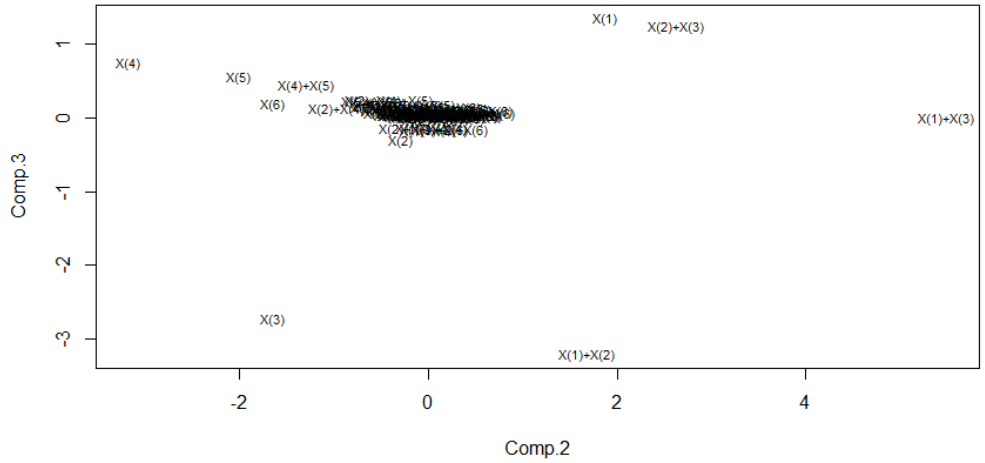

Рис. 2.13. График значений Comp.2 и Comp.3.

Аналогичные рассуждения для симптомов, состоящих из Y.

Таблица 2.33. Матрица факторных нагрузок для симптомов Y_k, $k = 2$.

	Comp.1	Comp.2	Comp.3	Comp.4
entr	0.1045669	0.59583156	0.78718729	-0.107138484
means_J	-0.1741724	0.56795112	-0.49147215	-0.635885865
freq_X.1.	-0.3916034	-0.06332506	0.08777376	-0.028257735
freq_X.2.	-0.3916034	-0.06332506	0.08777376	-0.028257735
freq_X.3.	-0.3916034	-0.06332506	0.08777376	-0.028257735
freq_X.4.	-0.3909877	-0.06225139	0.07968581	-0.034666922
freq_X.5.	-0.3890030	-0.06253003	0.15602310	-0.027755346
freq_X.6.	-0.3909198	-0.06412286	0.07752253	0.004323177
freq.X_without.X	-0.2041005	0.54634157	-0.28102637	0.761435597

По таблице 2.33 видно, что нужно строить график по 2 и 3 компонентам. Чем больше значение Comp.3, тем меньше means_J, freq.X_without.X и больше entr. Берем симптомы, попадающие в правую часть получившегося графика на рис. 2.14. Образуют устойчивые решения: Y(1) и Y(1)+Y(4).

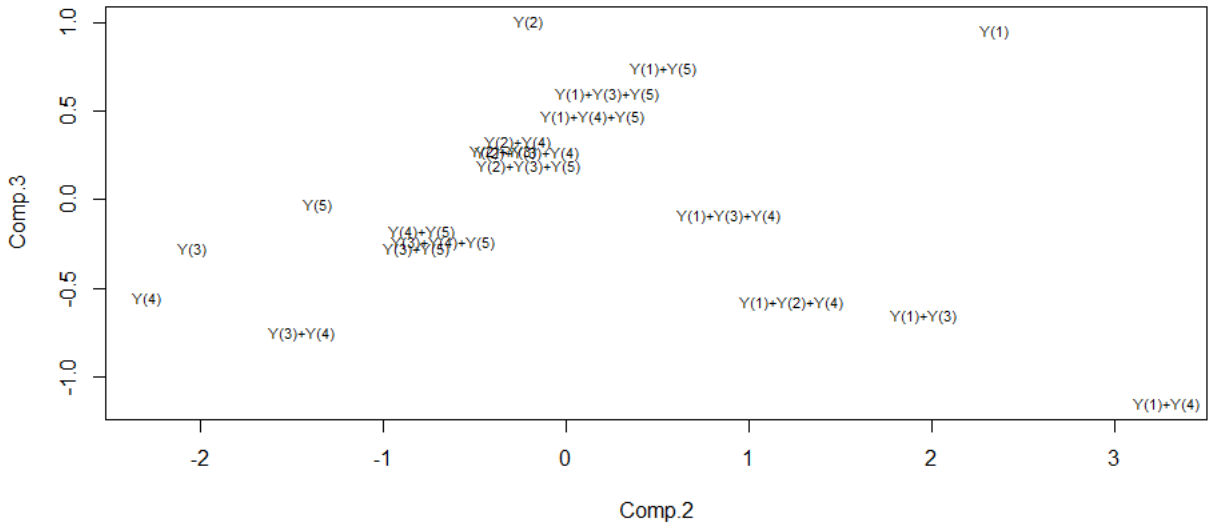

Рис. 2.14. График значений Comp.2 и Comp.3.

Были получены и упорядочены наиболее связанные сочетания X_k и Y_k $k = 1$, где в качестве меры зависимости рассматривается коэффициент неопределенности. Среди этих решений выделены устойчивые (табл. 2.34):

Таблица 2.34. Наиболее связанные подмножества X_k и Y_k, $k = 1$.

	names	J
1	X(1) - Y(1)	33.5573215377973
2	X(2)+X(3) - Y(1)	33.5573215377973
3	X(1) - Y(1)+Y(3)+Y(4)	31.9796195202503
4	X(2)+X(3) - Y(1)+Y(3)+Y(4)	31.9796195202503
5	X(1) - Y(1)+Y(3)	31.4577523721045
6	X(2)+X(3) - Y(1)+Y(3)	31.4577523721045
7	X(1) - Y(1)+Y(4)	28.5745578918406
8	X(2)+X(3) - Y(1)+Y(4)	28.5745578918406
9	X(2) - Y(1)	27.9763405325978
10	X(1)+X(3) - Y(1)	27.9763405325978
11	X(2) - Y(1)+Y(3)	26.9612320490325
12	X(1)+X(3) - Y(1)+Y(3)	26.9612320490325
13	X(2)+X(3) - Y(1)+Y(2)+Y(4)	25.9137966145166
14	X(1) - Y(1)+Y(2)+Y(4)	25.9137966145165
15	X(1) - Y(1)+Y(2)+Y(3)	25.1362342642408
16	X(2)+X(3) - Y(1)+Y(2)+Y(3)	25.1362342642408

2.3. Заключение

Таким образом, были получены следующие результаты:

Для множества признаков «До» и одного признака «После»:

1. Проведен факторный анализ для секции, содержащей оценки экспертов. Получилось, что эксперты оценивают достаточно адекватно.
2. Итог (выжил проект или нет) не прогнозируется экспертами.
3. Получены признаки блока информации «До», каждый из которых (в одиночку) оказывает влияние на итог:
   AGE,
   $WEB.SITE$,
   $INDUSTRY$,
   $FAMILY$,
   $TEACHER$,
   $KINDERGARTEN$.
4. Выделены подмножества признаков «До», наиболее связанные с выживаемостью проектов.
5. Выживают проекты, над которыми работали от 1 до 2 лет.
6. Найден номинативный представитель: признак AGE. При его добавлении к другим признакам, получаются подмножества, влияющие на итоговый признак.

Для множества признаков «До» и множества «После»:

1. Написана программа для оптимального поиска подмножеств признаков, основанная на алгоритме быстрого перечисления точек грассманиана.
2. Произведен канонический анализ. Получены наилучшие связи между подмножествами, в качестве меры зависимости рассматривается коэффициент неопределенности.
3. Сформулировано и доказано утверждение 1.

4. Реализован частотный способ поиска номинативных представителей.

5. Реализован метод поиска номинативных представителей, основанный на удалении признаков.

6. Получены номинативные представители обоих множеств X и Y при помощи методов многомерной статистики (табл. 2.35 и табл. 2.36):

   X(1) —$AGE1$ (до 2 лет работают над проектом),
   X(2)+X(3) — взаимодействие $AGE2$ и $AGE3$ (от 2 лет работают над проектом),
   Y(1) — $EVENTS.PUBLICATION$ (Публикации в СМИ о проекте),
   Y(1)+Y(4) — взаимодействие $EVENTS.PUBLICATION$ и $EVENTS.INVESTMENT$ (Публикации в СМИ о проекте + привлечены инвестиции).

Таблица 2.35. Таблицы сопряженности номинативных представителей двух множеств X и Y, Chi-square: $p.value$<2.2e-16.

	X(1)	Y(1) 0	Y(1) 1	Row Totals
Count	0	33	60	93
Column Percent		7,07%	70,59%	
Row Percent		35,48%	64,52%	
Count	1	434	25	459
Column Percent		92,93%	29,41%	
Row Percent		94,55%	5,45%	
Count	All Grps	467	85	552

	X(1)	Y(1)+Y(4) '0 0'	Y(1)+Y(4) '1 0'	Y(1)+Y(4) '1 1'	Y(1)+Y(4) '0 1'	Row Totals
Count	0	31	50	10	2	93
Column Percent		6,70%	69,44%	76,92%	50,00%	
Row Percent		33,33%	53,76%	10,75%	2,15%	
Count	1	432	22	3	2	459
Column Percent		93,30%	30,56%	23,08%	50,00%	
Row Percent		94,12%	4,79%	0,65%	0,44%	
Count	All Grps	463	72	13	4	552

Таблица 2.36. Таблицы сопряженности номинативных представителей двух множеств X и Y, Chi-square: $p.value$<2.2e-16.

	X(2)+X(3)	Y(1) 0	Y(1) 1	Row Totals
Count	' 0 0 '	434	25	459
Column Percent		92,93%	29,41%	
Row Percent		94,55%	5,45%	
Count	' 1 0 '	21	49	70
Column Percent		4,50%	57,65%	
Row Percent		30,00%	70,00%	
Count	' 0 1 '	12	11	23
Column Percent		2,57%	12,94%	
Row Percent		52,17%	47,83%	
Count	All Grps	467	85	552

	X(2)+X(3)	Y(1)+Y(4) ' 0 0 '	Y(1)+Y(4) ' 1 0 '	Y(1)+Y(4) ' 1 1 '	Y(1)+Y(4) ' 0 1 '	Row Totals
Count	' 0 0 '	432	22	3	2	459
Column Percent		93,30%	30,56%	23,08%	50,00%	
Row Percent		94,12%	4,79%	0,65%	0,44%	
Count	' 1 0 '	21	40	9	0	70
Column Percent		4,54%	55,56%	69,23%	0,00%	
Row Percent		30,00%	57,14%	12,86%	0,00%	
Count	' 0 1 '	10	10	1	2	23
Column Percent		2,16%	13,89%	7,69%	50,00%	
Row Percent		43,48%	43,48%	4,35%	8,70%	
Count	All Grps	463	72	13	4	552

В дальнейшем планируется:

1. Изучить значимость включения компонент в симптом.

Литература

1. Алексеева Н. П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2012. — 184 с.

2. Ананьевская П. В. Исследование конечно-линейных статистических моделей. Оптимизация и избыточность : дис. на соискание ученой степени кандидата физико-математических наук / П. В. Ананьевская ; Санкт-Петербургский гос. университет. — Санкт-Петербург, 2013. — 142 с.

3. Рао С. Р. Линейные статистические методы и их применение. — М. : Наука, 1968.

4. Алексеева Н. П. Учебное пособие по прикладной статистике. Часть 2. Многомерные методы. — Санкт-Петербург, 2014.

5. Ермаков М. С., Сизова А. Ф., Товстик Т. М. Учебное пособие: Элементы математической статистики. — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2001. — 148 с.

6. Воробьев О. Ю. Эвентология. — Красноярск : Сиб. фед. ун-т, 2007.