Канонический анализ категориальных данных с приложением в маркетинге

Григорьева Ирина Владимировна, гр. 422

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

Научный руководитель:к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: исследователь, ВШЭ Смирнов И.Б.

Санкт-Петербург, 2016г.

Постановка задачи

Задача:
Исследование зависимости между двумя наборами номинальных признаков $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ по аналогии с каноническим корреляционным анализом.

Мера зависимости:
Информационный коэффициент неопределенности.

Проблема:

1. Расширения X и Y на основе операций над полем $\mathbb{F}_2$.
2. Отбор наиболее информативных подпространств.
3. Критерий отсева компонент подпространств.

Прикладная задача

Имеются данные «Инновации в образовании»(Центр исследований инноваций в образовании), состоящие из трех блоков:

1. Заявки на конкурс ($N = 552$).
2. Оценки экспертов к каждой из заявок:
   - Общая оценка эксперта.
   - Новизна и оригинальность идеи.
   - Актуальность решаемых проблем.
   - Целесообразность используемых механизмов.
   - Возможность тиражирования.
3. Итоговые характеристики: анкета, которую участники заполняли через год.

Прикладная задача: Возможность прогнозирования развития проекта через год по данным заявкам и оценкам экспертов.

Используемые методы

1. Факторный анализ.
   - Редукция размерности в исследовании оценок экспертов для количественных признаков.
   - Отбор наиболее информативных компонент.
2. Дисперсионный анализ.
   - Качество оценивания выживаемости экспертами.
3. Симптомно-синдромальный анализ категориальных данных на основе конечных геометрий.
4. Алгоритм быстрого перечисления точек грассманиана Ананьевской П.В.

Интерпретация фактора: актуальность решаемых проблем и целесообразность используемых механизмов.

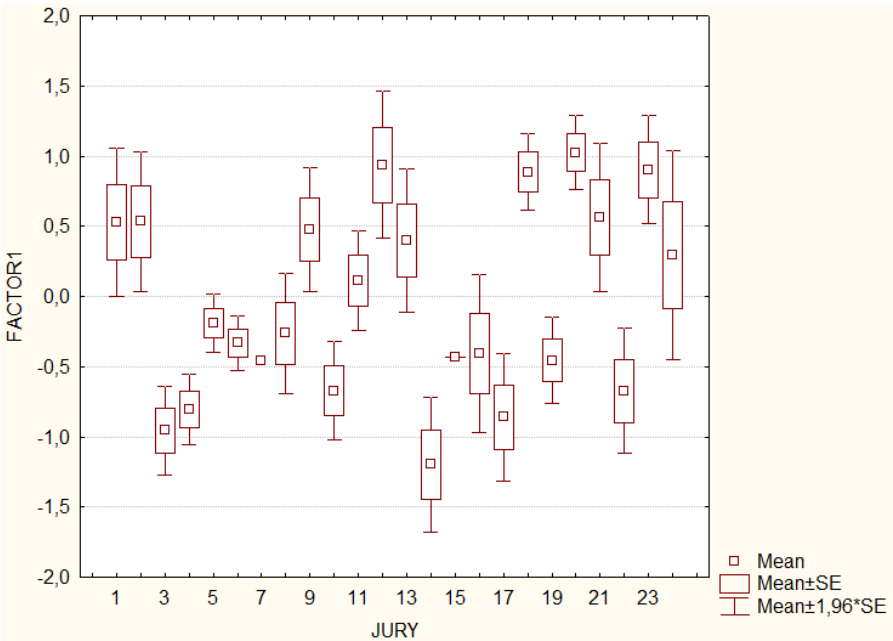


Рис.: Диаграмма размаха ($Factor1$ и $JURY$).

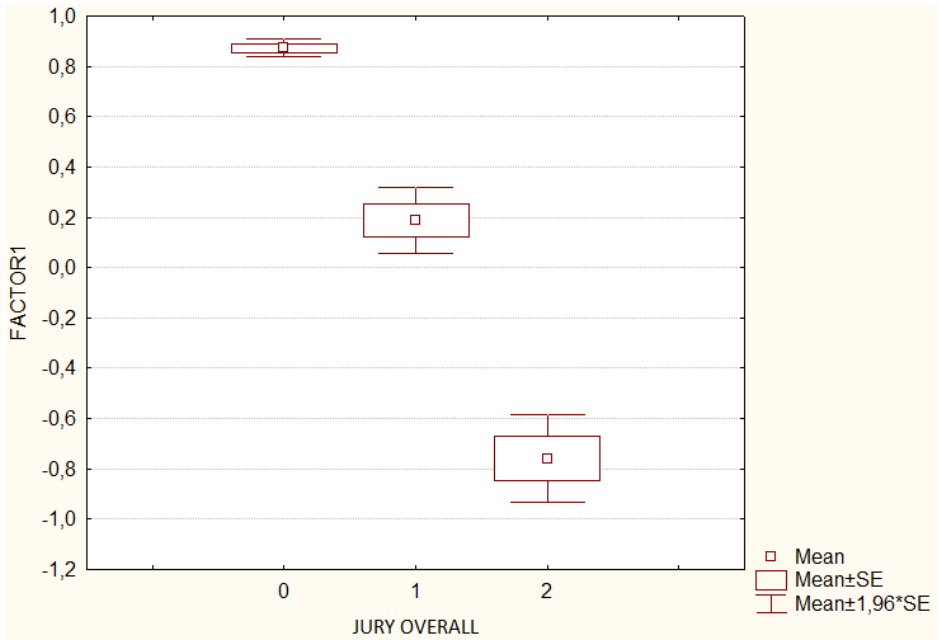
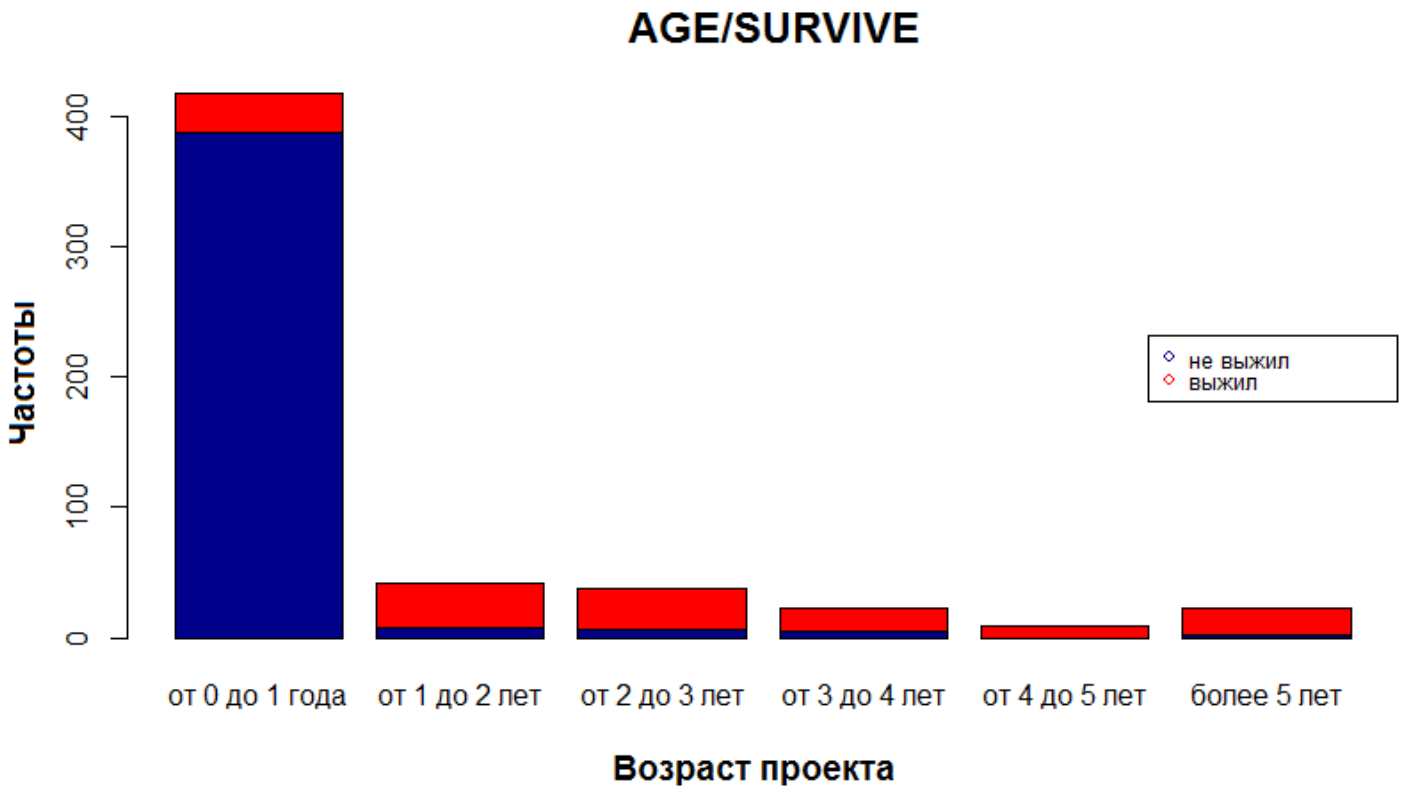


Рис.: Диаграмма размаха ($Factor1$ и общая оценка).

Оценивание выживаемости проектов

- Прогнозируемость по оценкам перспективности:
  С помощью статистики хи-квадрат проверена гипотеза о независимости признаков (оценка эксперта и $SURVIVE$ – продолжают ли работу над проектом).
- Прогнозируемость по категориальным данным заявки.



Основные определения

Симптом – линейная комбинация дихотомических признаков X_i вида $\sum\limits_{i=1}^{m} a_i X_i (\mathrm{mod}\, q)$, $a_i \in \mathbb{F}_q$.

Ранг симптома – количество $a_i : a_i \neq 0, i = \overline{1, m}$.

Синдром k-го порядка – совокупность л. н. симптомов $x_0, \ldots, x_k$ вида $\sum\limits_{i=0}^{k} \beta_i x_i (\mathrm{mod}\, q)$, где $\beta_i \in \mathbb{F}_q$ не равны нулю одновременно.

Обозначение: $\mathbf{X_k} = \langle x_0, \ldots, x_k \rangle$.

Номинативный представитель – симптом наименьшего ранга, при исключении которого значимо изменяются свойства синдрома.

Замечание:

Синдром k-го порядка – это проективная геометрия $\mathrm{PG}(k, q)$, симптомы – элементы проективной геометрии.

Основные определения

Энтропия с.в $\xi = \begin{pmatrix} x_1 & \cdots & x_m \\ p_1 & \cdots & p_m \end{pmatrix}$ $\mathrm{H}(\xi) = -\sum_{i=1}^{m} p_i \log_2 p_i.$

Совместная информация с.в ξ и η

$$\mathrm{I}(\xi, \eta) = \mathrm{H}(\xi) + \mathrm{H}(\eta) - \mathrm{H}(\xi, \eta).$$

Односторонний коэффициент неопределенности между с.в ξ и η

$$\mathrm{J}(\xi|\eta) = \frac{\mathrm{I}(\xi, \eta)}{\mathrm{H}(\eta)}.$$

Отбор наиболее связанных подпространств

$X_1, \ldots, X_6$ – заявки, $Y_1, \ldots, Y_5$ – анкеты.

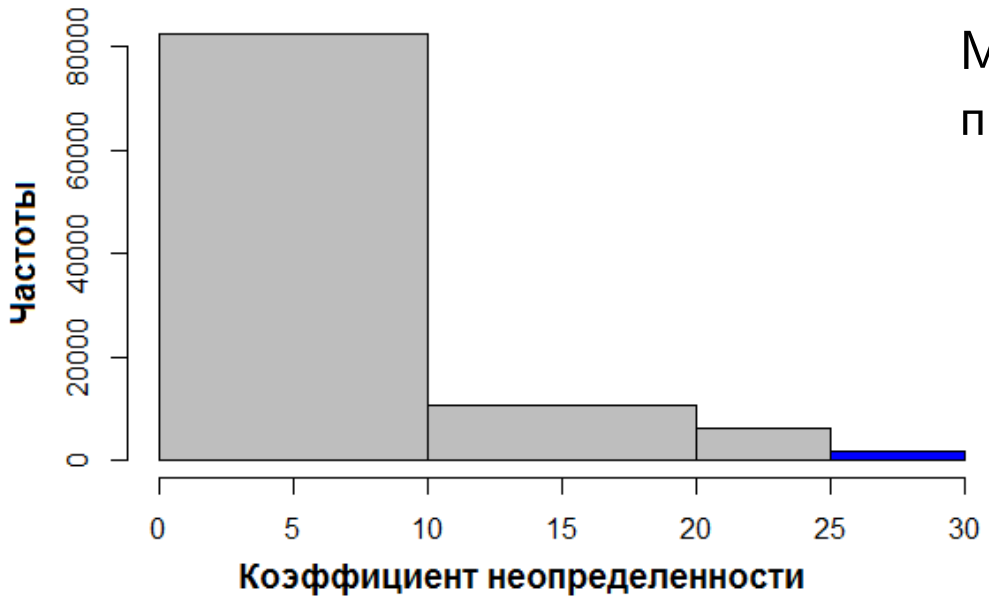
Выделены $\mathbf{X_1}$, $\mathbf{Y_1}$ по алгоритму быстрого перечисления точек грассманиана.

$\mathbf{X_1} = \langle x_{\delta_1}, x_{\delta_2} \rangle$: $\#\mathbf{X_1} = 651$

$\mathbf{Y_1} = \langle y_{\nu_1}, y_{\nu_2} \rangle$: $\#\mathbf{Y_1} = 155$

$\#\,\mathrm{J} = 100905$

$\#\{\mathrm{J}(\mathbf{X_1}|\mathbf{Y_1}) : \mathrm{J}(\mathbf{X_1}|\mathbf{Y_1}) > 25\%\} = 740$



Методы поиска номинативных представителей:

1. Селективный.
2. Частотный.
3. Комбинированный на основе главных компонент.

Селективный метод поиска номинативных представителей

Утверждение

Пусть $\mathbf{X} = \langle x_1, \ldots, x_m \rangle$, $\mathbf{Y} = \langle y_1, \ldots, y_n \rangle$, $x \in \mathbf{X}$,
$\mathrm{J}_L = \mathrm{J}(\mathbf{X}|\mathbf{Y}) - \mathrm{J}(\mathbf{X}\backslash x|\mathbf{Y})$, $\mathrm{J}_R = \mathrm{J}(\mathbf{Y}|\mathbf{X}) - \mathrm{J}(\mathbf{Y}|\mathbf{X}\backslash x)$.

Если $\mathrm{H}(\mathbf{X}) = \sum_{i=1}^{m} \mathrm{H}(x_i)$, $\mathrm{H}(\mathbf{Y}) = \sum_{j=1}^{n} \mathrm{H}(y_j)$, *то* $\mathrm{J}_L = \mathrm{J}_R = 0$.

Если $\mathbf{X}$ *и* $\mathbf{Y}$ *независимые в совокупности, то* $\mathrm{J}_L > 0$, $\mathrm{J}_R \neq 0$.

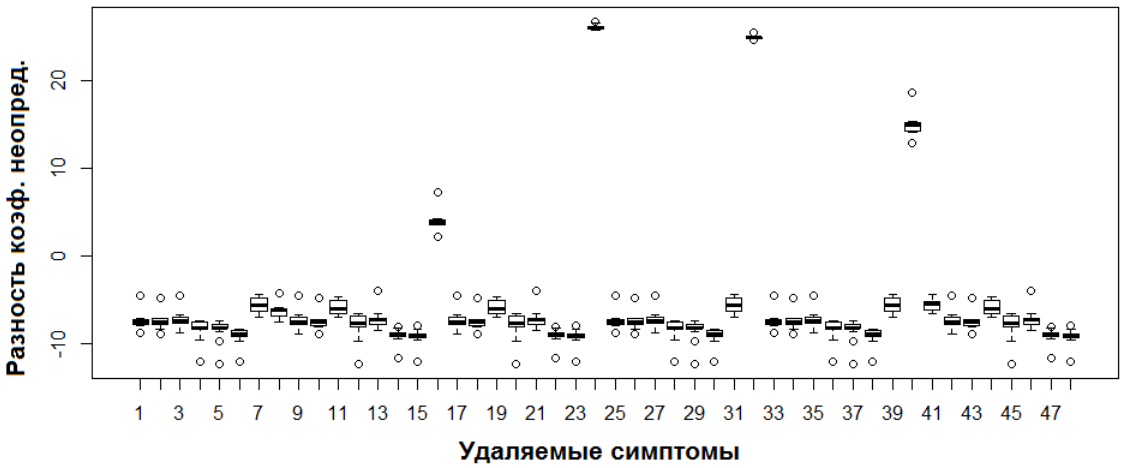
Критерии: критерий знаков, ранговый критерий Вилкоксона для зависимых выборок.

Результат:

- набор $\mathrm{J}_2 = \mathrm{J}(y|\mathbf{X_0})$ статистически значимо уменьшается по сравнению с исходным $\mathrm{J}_1 = \mathrm{J}(y|\langle x, \mathbf{X_0} \rangle)$
  при удалении $x \in \{x_1, x_2, x_2 + x_3, x_1 + x_3\}$ над полем $\mathbb{F}_2$.
- набор $\mathrm{J}_2 = \mathrm{J}(x|\mathbf{Y_0})$ статистически значимо уменьшается по сравнению с исходным $\mathrm{J}_1 = \mathrm{J}(x|\langle y, \mathbf{Y_0} \rangle)$ при удалении
  $y \in \{y_1, y_1 + y_3, y_1 + y_3 + y_4, y_1 + y_4, y_1 + y_2 + y_4\}$ над полем $\mathbb{F}_2$.

Иллюстрация селективного метода

Диаграммы размахов для симптомов наиболее связанных подпространств.

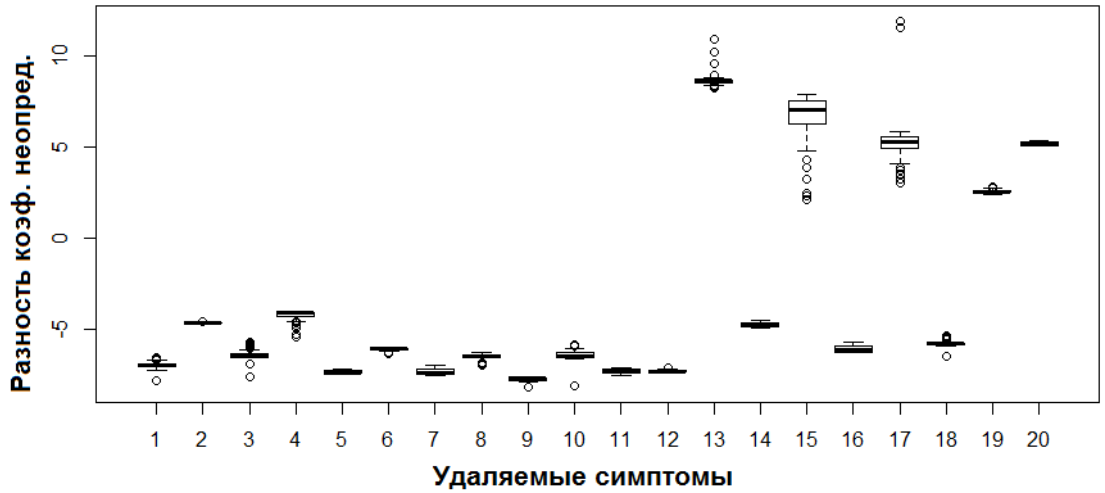


24 — $x_2 + x_3 \pmod 2$,

32 — x_1,

40 — $x_1 + x_3 \pmod 2$,

16 — x_2.



13 — y_1,

15 — $y_1 + y_4 \pmod 2$,

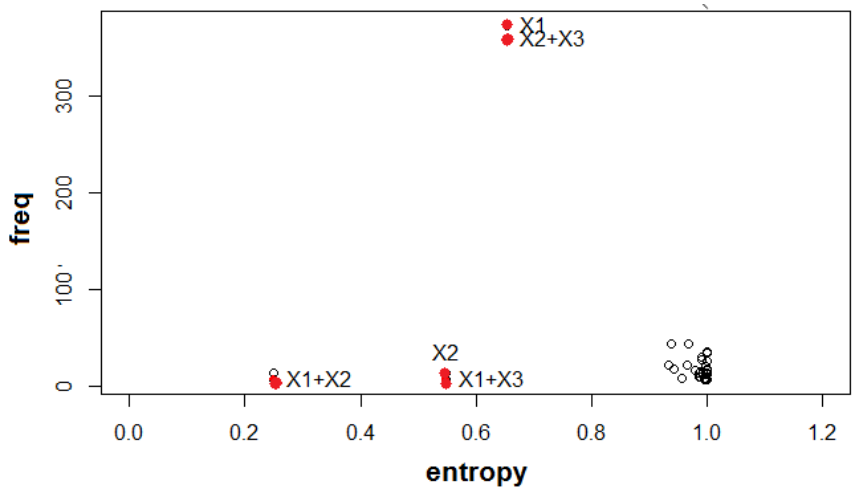
17 — $y_1 + y_3 \pmod 2$,

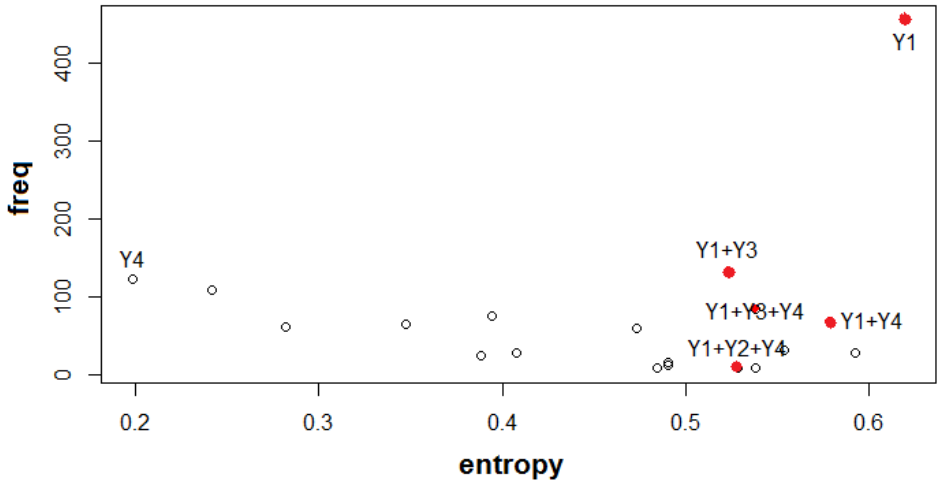
20 — $y_1 + y_2 + y_4 \pmod 2$,

19 — $y_1 + y_3 + y_4 \pmod 2$.

Частотный метод поиска номинативных представителей

Красным цветом на графиках отмечены симптомы, найденные предыдущим способом.





$x_1, \ldots, x_m \in X, y_1, \ldots, y_n \in Y$, $l = 25$.

$$\text{freq}_x = \#\{\mathbf{X_1} : x \in X \subset \mathbf{X_1},\ \text{J}(\mathbf{X_1}|\mathbf{Y_1}) > l\ \forall \mathbf{Y_1}\}$$
$$\text{entropy}_x = \text{H}(x)$$

$$\text{freq}_y = \#\{\mathbf{Y_1} : y \in Y \subset \mathbf{Y_1},\ \text{J}(\mathbf{X_1}|\mathbf{Y_1}) > l\ \forall \mathbf{X_1}\}$$
$$\text{entropy}_y = \text{H}(y)$$

Комбинированный метод поиска номинативных представителей

Интерпретация для X:

Comp.2 – информационная значимость симптомов для канонической корреляции.

Comp.3 – информативность симптомов.

Номинативные представители:

x_1 и $x_2 + x_3 \pmod 2$.

Интерпретация для Y:

Comp.2 – информационная значимость наиболее информативных симптомов.

Comp.3 – информативность симптомов при информационной незначимости.

Номинативные представители:

y_1 и $y_1 + y_4 \pmod 2$.

Решение прикладной задачи

Получены номинативные представители обоих множеств:

Для множества X:

x_1 – $AGE1$ (до 2 лет работают над проектом),
$x_2 + x_3 \pmod 2$ – взаимодействие $AGE2$ и $AGE3$ (от 2 лет работают над проектом).

Для множества Y:

y_1 – публикации в СМИ о проекте,
$y_1 + y_4 \pmod 2$ – фактор успешности (либо публикации в СМИ о проекте, либо привлечены инвестиции).

Результаты

- Эксперты оценивают адекватно.
- Итог не прогнозируется экспертами.
- Хуже всего выживают проекты над которыми работали меньше года.

- Написана программа на языке R для оптимального поиска подмножеств признаков, основанная на алгоритме быстрого перечисления точек грассманиана.
- Произведен канонический анализ. Выделены наиболее связанные подмножества с использованием коэффициента неопределенности.
- Получены номинативные представители обоих множеств.

Основной результат и планы на будущее

Основной результат:

- Доказано утверждение о знаке разности коэффициентов неопределенности при удалении одного симптома из синдрома в случае независимости синдромов в совокупности.
- Разработаны методы поиска номинативных представителей на примере практической задачи:
  - Селективный метод.
  - Частотный метод.
  - Комбинированный метод на основе главных компонент.

В дальнейшем планируется:

- Разработка критерия информативности компонент симптома.
- Расширение X и Y на основе операции умножения над $\mathbb{F}_2$.