Статистический анализ бинарных данных с приложением в иммунологии

Соколиков Евгений Алексеевич, гр. 15.Б04-мм

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

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: к.ф.-м.н., младший научный сотрудник Ананьевская П.В.

Санкт-Петербург
2019г.

Эксперимент: изучение адаптивной способности индивидуумов в экстремальных условиях.

- 26 индивидов, каждому соответствует бинарный временной ряд длины 254.
- Единица кодирует наличие цитокиновой реакции, ноль — ее отсутствие.

Задачи

- Изучить методы исследования бинарных данных.
- Проанализировать исследуемые данные.

Методы:

- Критерий серий — проверка рядов на стационарность.
- Приоритетные кривые — визуализация изменения вероятности появления единицы на протяжении эксперимента.
- Суперсимптомный анализ — поиск повторяющегося из ряда в ряд фрагмента.
- Кластеризация - использование результатов двух предыдущих методов для разбиения индивидов на группы.

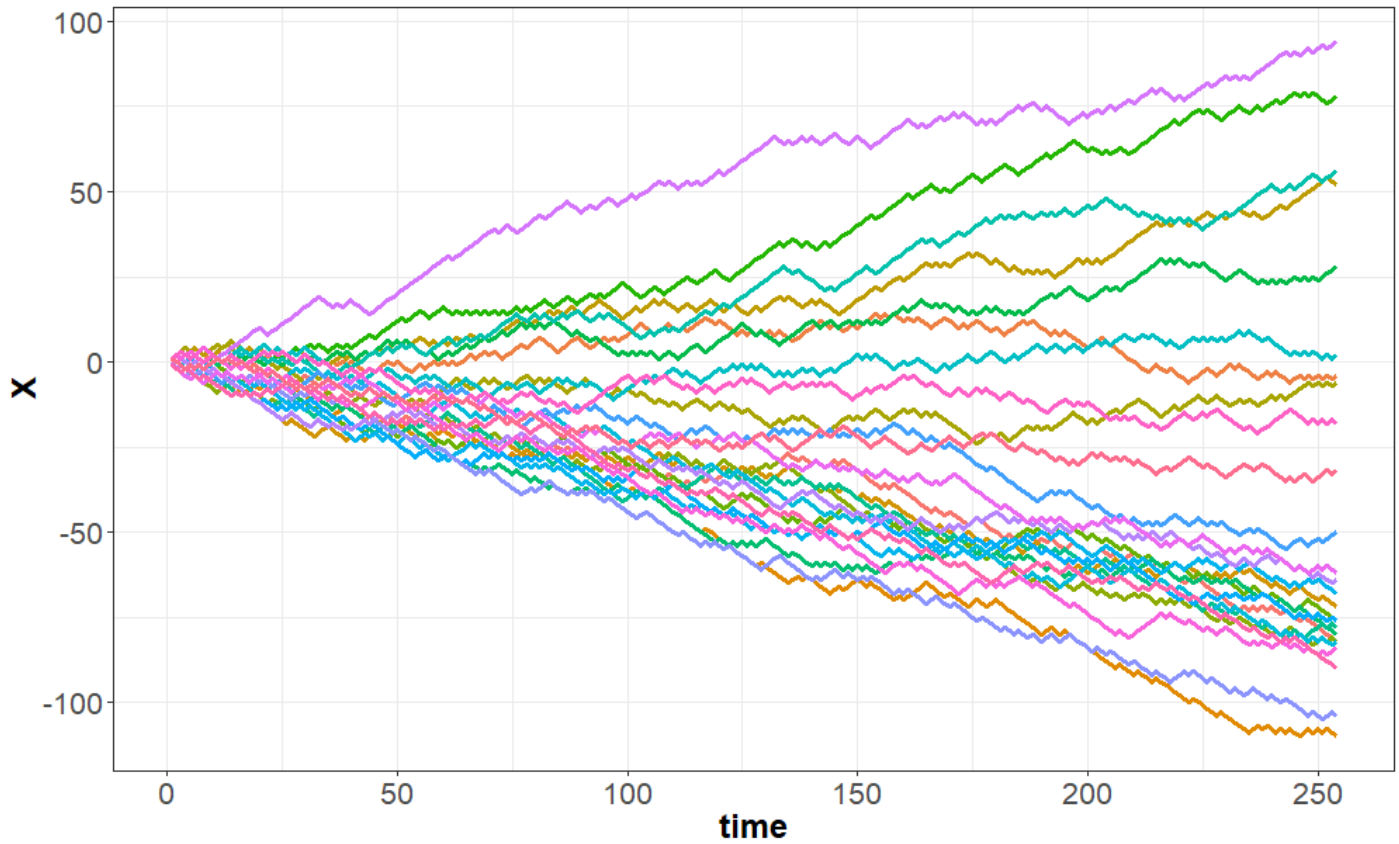
Приоритетные кривые [Alexeeva N., 2012]

- Приоритетными кривыми будем называть кумулятивные ряды вида

$$\tilde{\mathbb{X}}(t) = \sum_{j=1}^{t} (2\mathbb{X}(j) - 1)$$

  представляющие собой реализации случайных процессов.
- Для каждого индивида можно построить приоритетную кривую, которая будет возрастать, когда наблюдается цитокиновая реакция, и убывать в противном случае.
- Линейная форма приоритетной кривой будет указывать на примерно постоянную вероятность появления единицы. Выпуклость-вогнутость — на ее перемену.

Приоритетные кривые. Построение



Приоритетные кривые индивидов.

Вывод: Почти линейная форма приоритетных кривых говорит о примерно постоянной вероятности возникновения цитокиновой реакции у каждого индивида. Это совпадает с результатами, полученными при применении критерия серий.

- Пусть $\mathbb{X} = (X_1, \ldots, X_m)$ — бинарная матрица.
- Испульсно-упорядоченный вектор произведений:

$$\mathrm{V}_1 = X_1,\ \mathrm{V}_j = (\mathrm{V}_{j-1}, X_j, X_j \cdot \mathrm{V}_{j-1}).$$

- Суперсимптомы:

$$Y_\tau(\mathbb{X}) = \alpha^{\mathrm{T}} \mathrm{V}_m \pmod 2 = \sum_{j=1}^{2^m-1} a_j \prod_{i=1}^{m} X_i^{b_i} \pmod 2,$$

где $\alpha^{\mathrm{T}} = (a_1, \ldots, a_{2^m-1})$, $a_j, b_i \in \{0, 1\}$, $\tau = \{j : a_j = 1\}$.

- Импульсный ранг: $R(\tau) = \sum_{q=1}^{j} 2^{t_q - 1}$ — порядковый номер суперсимптома Y_τ в импульсно-упорядоченном ряду при $\tau = \{t_1, \ldots, t_j\}$.

- Ω_m - множество всевозможных бинарных векторов $\mathrm{X} = (x_1, \ldots, x_m)$ длины m.
- $\alpha^{\mathrm{T}} = (a_1, \ldots, a_{2^m-1})$ из $Y_\tau(\mathrm{X}) = \alpha^{\mathrm{T}} \mathrm{V}_m \pmod 2$
- $k(\mathrm{X}) = \sum_{j=1}^{m} 2^{j-1} x_j$ — нумерация элементов Ω_m.
- $\Theta_\tau = \{k(\mathrm{X}) : Y_\tau(\mathrm{X}) = 1,\ \mathrm{X} \in \Omega_m\}$
- $\beta_\tau = (\beta_1, \ldots, \beta_{2^m-1})$, где $\beta_i = 1$, если $i \in \Theta_\tau$, и $\beta_i = 0$ иначе.

Теорема

Пусть $G_1 = 1$, $\mathbb{O}_{l,m}$ — нулевая матрица размерности $l \times m$, а $\mathbf{1}_k$ — вектор, состоящий из одних единиц,

$$G_{k+1} = \begin{pmatrix} G_k & \mathbb{O}_{n,1} & \mathbb{O}_{n,n} \\ \mathbb{O}_{1,n} & 1 & \mathbb{O}_{1,n} \\ G_k & \mathbf{1}_n & G_k \end{pmatrix}.$$

Тогда

- $G_m\ \alpha \pmod 2 = \beta_\tau$ *для любого симптома* Y_τ.
- $\forall j\ G_j^2 \pmod 2 = \mathbb{I}$, *где* $\mathbb{I}$ *- матрица соответствующей размерности с единицами на главной диагонали.*

Суперсимптомный анализ. Показатель рентабельности [Alexeeva N., 2012]

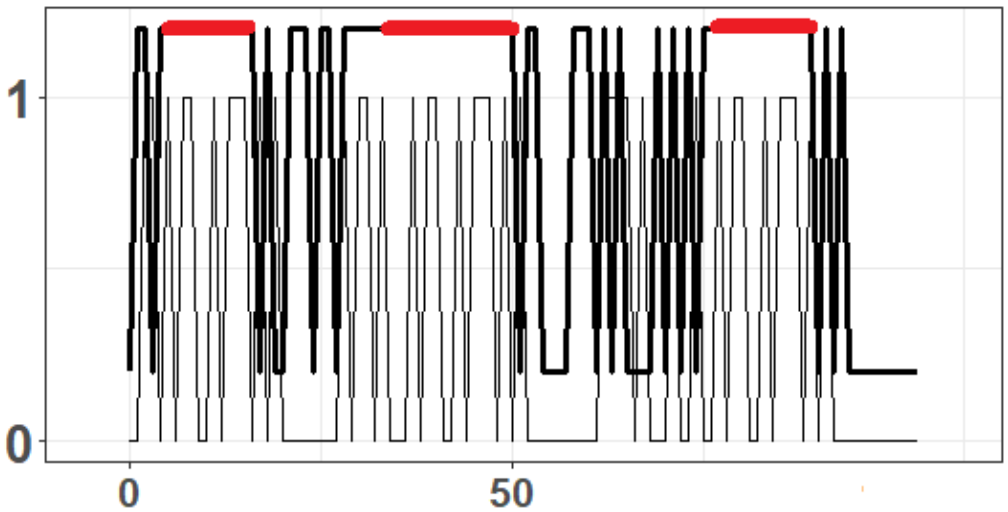
- Наибольшее число единиц.
- Наименьшее количество переключений от одного знака к другому.

Определение

Показателем рентабельности $Ren(R(\tau))$ *будем называть объединение этих характеристик:*

$$Ren(R(\tau)) = \frac{Y_\tau^T e_L}{(\Theta Y_\tau (mod\ 2))^T e_{L-1}},$$

где Θ *— матрица размерности* $L-1$ *на* L*, у которой* $\Theta_{i,i} = \Theta_{i,i+1} = 1$*, остальные нули.*



- Объединяем все ряды в один категориальный ряд. Будем рассматривать его траекторную матрицу $X = [X_1, \ldots, X_m]$. Построим соответствующие ей суперсимптомы.
- В данном случае наиболее информативный суперсимптом — $Y_{1,2,3,4,5,6,7}$.
- По теореме: $\beta_{1,2,3,4,5,6,7} = (1, 1, 1, 1, 1, 1, 1)$
- Интерпретация:

$$Y_{1,2,3,4,5,6,7} = 1 - \overline{X_1} \cdot \overline{X_2} \cdot \overline{X_3}.$$

Итоги: данный суперсимптом идентифицирует периоды отсутствия цитокиновой реакции длины хотя бы три.

- Замечание: на реальных данных такой суперсимптом будет всегда иметь показатель рентабельности не меньше любого другого суперсимптома.

Определение

Альтернативным показателем рентабельности $Ren_2(R(\tau))$ *будем называть число обратное количеству переключений в суперсимптоме:*

$$Ren_2(R(\tau)) = \frac{1}{(\Theta Y_\tau (mod\ 2))^T e_{L-1}},$$

где Θ *— матрица размерности* $L-1$ *на* L*, у которой* $\Theta_{i,i} = \Theta_{i,i+1} = 1$*, остальные нули.*

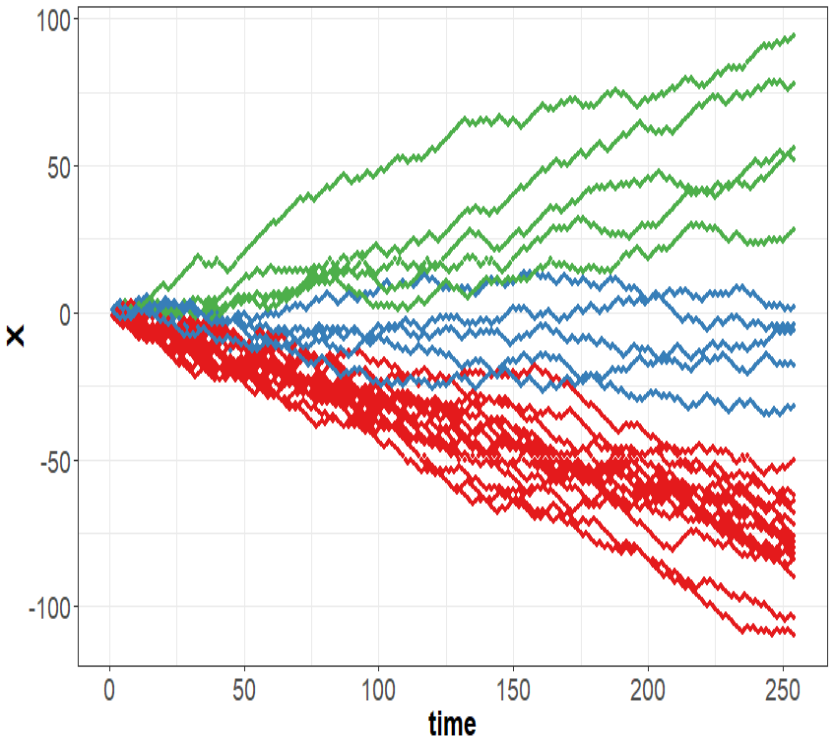
- По альтернативному показателю рентабельности наиболее информативный суперсимптом — Y_{64}.
- По теореме: $\beta_{64} = (0,0,0,0,0,0,1)$

Интерпретация: данный суперсимптом идентифицирует периоды наличия цитокиновой реакции длины хотя бы три.

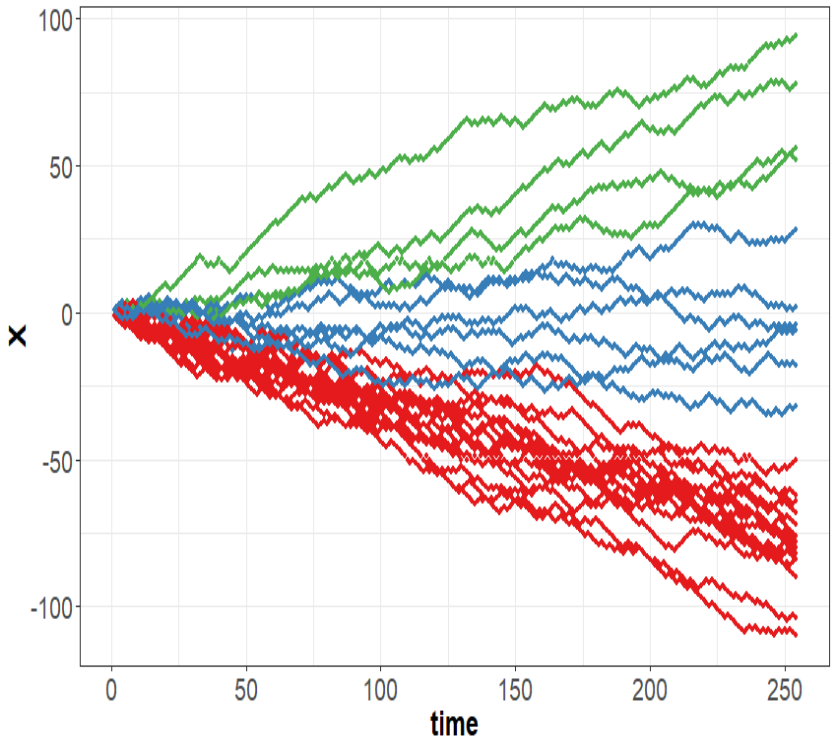
Другие критерии информативности. Второй

- Пусть S — вектор количества единиц в некотором суперсимптоме для каждого индивида, W — аналогичный вектор количества переключений.
- Упорядочим все суперсимптомы по первому собственому числу ковариационной матрицы вектора (S, W).
- Наилучший суперсимптом по такой метрике — Y_{52}.
- По теореме: $\beta_{52} = (0, 0, 1, 0, 1, 1, 1)$
- Интерпретация: данный суперсимптом накрывает фрагменты длины три, в которых наблюдаются хотя бы две цитокиновых реакции.

- Кластеризация проводится по двум параметрам — количеству единиц и количеству переключений в суперсимптоме у каждого индивида.



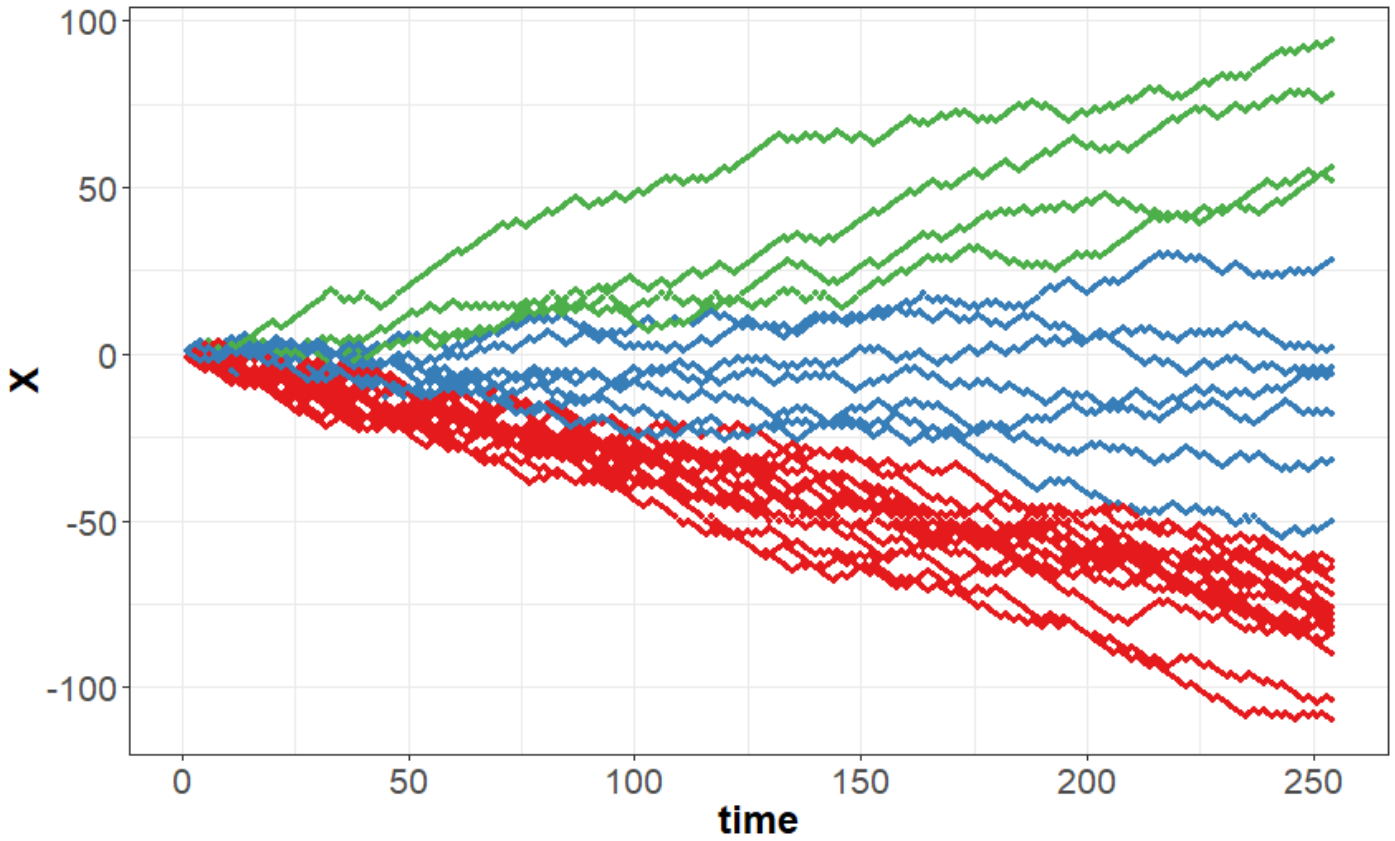
а) Y_{127}



б) Y_{64}

Кластеризация по наиболее информативным симптомам, полученным по первым двум показателям рентабельности

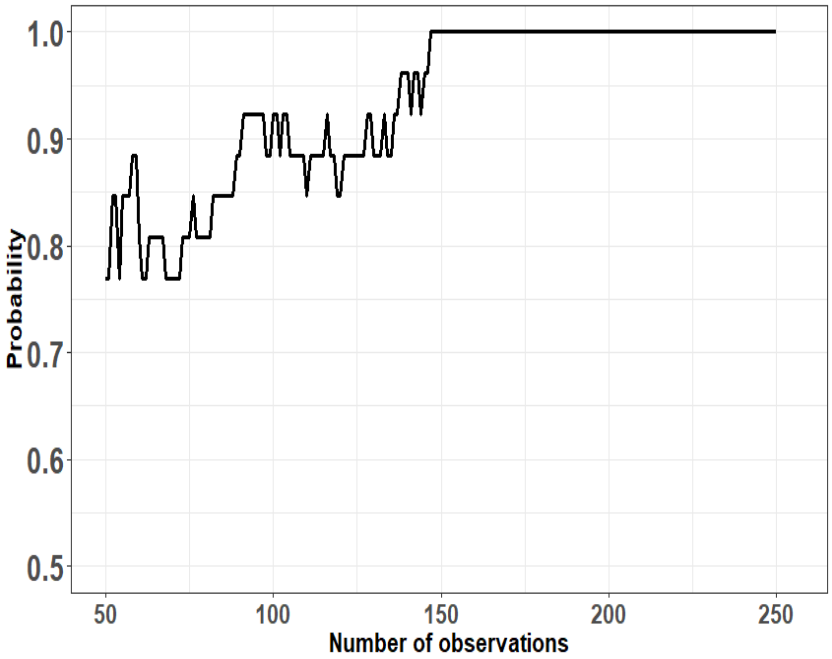
Наилучшая кластеризация



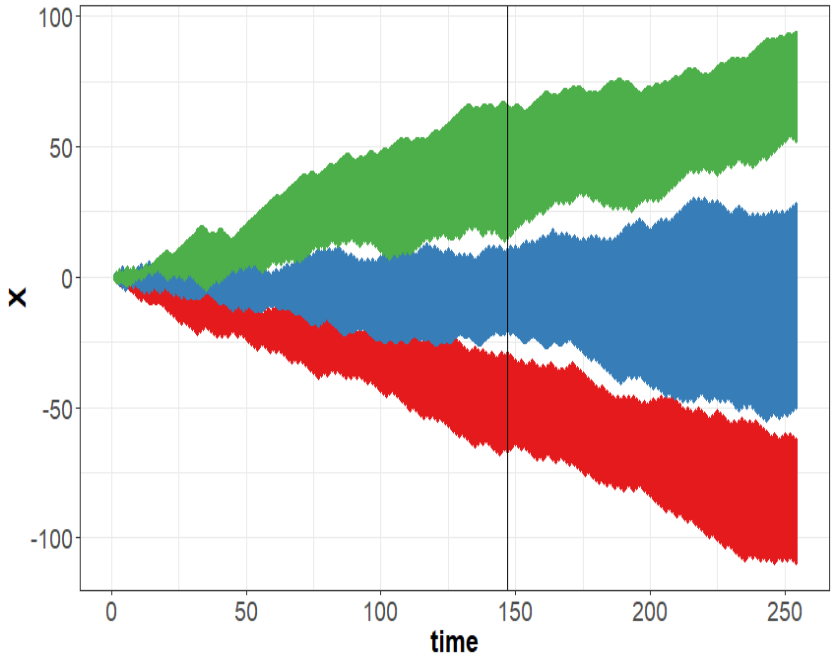
Кластеризация по 52 суперсимптому.

Вопрос: Можно ли сократить время проведения эксперимента, сохранив возможность правильной автоматической кластеризации?

Сокращение времени эксперимента



а) Вероятность правильной классификации



б) Кластеризация и наименьшее допустимое время

Вывод: Характеристики 52 суперсимптома позволяют автоматически правильно классифицировать индивидов, начиная с длины эксперимента в 147 наблюдений.

Результаты:

- Изучены методы статистического анализа бинарных данных.
- Доказана теорема о восстановлении вектора коэффициентов суперсимптома по веткору реализаций.
- Соответствующие методы применены к тестовым данным, исследуемым данным, проанализированы полученные результаты.
- Рассмотрены дополнительные критерии информативности суперсимптомов, более адекватные для рядов с большей энтропией.
- Получена автоматическая классификация индивидов по их адаптивным способностям для наименьшей длительности эксперимента.
- Разработано соответствующее программное обеспечение на R.