Федеральное государственное автономное
образовательное учреждение высшего
образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии

Базовая кафедра медико-биологических систем и комплексов

УТВЕРЖДАЮ
Заведующий кафедрой
_____ А.Н. Шуваев
« 22» июня 2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Сравнения алгоритмов классификации для определения положения изолированных пациентов

03.04.02 Физика
03.04.02.09 Технологическое сопровождение ядерной медицины и медицинского оборудования

Научный руководитель _____22.06.2020г. к.ф-м.н. А.Н. Шуваев

Выпускник _____22.06.2020г. А.Ю. Чмурин

Рецензент _____22.06.2020г. к.б.н., доцент Л.В. Степанова

Красноярск 2020

РЕФЕРАТ

Магистерская диссертация по теме «Сравнения алгоритмов классификации для определения положения изолированных пациентов» содержит 44 страниц текстового документа, 37 использованных источников, 15 рисунков, 3 таблицы, 7 формул, 1 приложение.

КЛАССИФИКАЦИЯ, АЛГОРИТМ K-БЛИЖАЙШИХ СОСЕДЕЙ, НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР, СЛУЧАЙНЫЙ ЛЕС, RFID, ПАДЕНИЯ, ПОЛОЖЕНИЯ.

Объект исследования – база данных изучения движения пожилых людей.

Работа посвящена сравнению алгоритмов классификации (k-ближайших соседей, Наивный байесовский классификатор, Случайный лес). Это необходимо для выбора лучшего классификатора для данных с носимых пассивных RFID меток. Актуальной задачей является предотвращение критических ситуаций, связанных с падениями изолированных пациентов.

В проведенном исследовании были проанализированы вышеперечисленные классификаторы. По результатам анализа было проведено их сравнение и сделан вывод, что алгоритм k-ближайших соседей является лучшим для задачи классификации определения положения изолированных пациентов.

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ

CRFID (англ. **R**adio **F**requency **ID**entification) – датчик компьютерной радиочастотной идентификации

ОП – обучающая подвыборка

СКО – сумма квадратов отклонений

ВВЕДЕНИЕ

На сегодняшний день в современном мире постоянно развивающихся технологий также совершенствуется и медицина. Однако, к сожалению, из-за растущего населения и улучшения качества медицинских услуг непрерывно растет количество людей, требующих к себе повышенного внимания, находясь в медицинских учреждениях. Будь это человек после операции, находящийся на реабилитации или просто человек преклонного возраста. И тому, и другому необходим постоянный мониторинг, чтобы избежать падений. Считается, что падения среди стационарных больных являются обычным явлением. В исследовании Хитчо и др. 2004 г. [14] приводится статистика падений людей. Согласно исследованию падения обычно происходят в больницах и домах престарелых. Также в литературных источниках сообщается, что падения в этих местах происходят в палатах пациента (85%) и во время передвижения. Кроме того, большинство падений происходит вокруг кровати и стула [32, 21]. На 1000 пациентов-дней приходится до 7 падений. Приблизительно треть падений в стационаре приводит к легким травмам, а до 6% - к серьёзным, вплоть до летального исхода. Данная статистика остро ставит вопрос непрерывного постоянного мониторинга изолированных пациентов.

Кроме того, падения значительно увеличивают срок пребывания пациентов в больницах, что увеличивает затраты на их лечение [13]. Также увеличенное время нахождения в больничном учреждении оставляет огромный негативный след на моральном состоянии не только самих пациентов, но и медицинский персонал и других людей, которые заботятся о них [20]. Постоянное отслеживание действий пациента и распознавание его деятельности дают возможность вмешаться и предотвратить падение или обеспечить повышенное внимание со стороны людей, осуществляющих уход [31, 29], в отличие от распознавания падений [22, 18]. Стратегия обнаружения падения не является стратегией снижения падений.

Существуют исследования, которые были сосредоточены на обнаружении выхода из кровати [5, 6, 12, 29, 31], они основывались на одном или нескольких датчиках, размещенных на или вокруг кровати. В большинстве из этих методов использовались датчики давления, обеспечивающие различные результаты производительности из-за использования нескольких типов чувствительных блоков. Кроме того, датчики давления были признаны ненадежными у пациентов весом менее 45,4 кг, что является средним весом для слабых пациентов, но в сочетании с другими датчиками была достигнута хорошая производительность [6]. Для расположения этих блоков были выбраны различные места: коврики для кроватей, рельсы для кроватей, коврики для пола. Такое расположение делает их подверженными постоянным механическим нагрузкам, требующим регулярного технического обслуживания и замены, то есть пристального внимания. Кроме того, эти устройства требуют тщательной очистки, поскольку они могут подвергаться воздействию жидкостей организма и других загрязняющих материалов.

Использование носимых батарейных датчиков для распознавания активности широко изучалось в исследованиях [1, 2, 7, 17, 19, 26, 27, 28, 35], однако, их применение для мониторинга пожилых людей было ограничено неудобством ношения этих устройств [33], а также минусом являются некоторые требования, например, такие как обслуживание аккумуляторов.

Возможность беспроводного питания нового класса датчиков, носимых на теле, на примере пассивных (без батарей) датчиков компьютерной радиочастотной идентификации (CRFID) [25], создает новые возможности для анализа движений человека. Пассивные датчики – удобные и простые в использовании устройства для распознавания активности [15], потому что они легкие, компактные и без батарей. Следовательно, такие датчики идеальны в качестве носимых датчиков для пожилых людей, где незаметность, удобство ношения и эргономические требования являются важными факторами для воплощения технологии на практике [10], так как планируется использовать эти

датчики на постоянной основе. Поэтому мы рассматриваем использование пассивного датчика CRFID на теле для отслеживания основных движений изолированных пациентов: 1) сидеть на кровати; 2) сидеть на стуле; 3) лежать; и 4) перемещаться, чтобы реализовать систему амбулаторного мониторинга.

Целью работы является выбор наиболее точного метода классификации для мониторинга передвижения изолированных пациентов.

Для достижения цели были поставлены следующие задачи исследования:

1. Из литературных источников выделить методы классификации, подходящие к данной задаче;
2. Выполнить классификацию выделенными методами;
3. Провести анализ начальных данных и подобрать лучшие параметры методов;
4. Сравнить точность методов и выбрать наиболее подходящий метод.

1 Литературный обзор

Классификация - это контролируемый подход машинного обучения, при котором алгоритм учится на основе предоставленных ему данных, а затем использует это обучение для классификации новых наблюдений.

Ни один из алгоритмов обучения не является лучшим для всех задач классификации [34]. Следовательно, в каждом случае соответствующий алгоритм должен быть выбран путем оценки его производительности.

1.1 Алгоритм k-ближайших соседей

Классификатор k-ближайших соседей является одной из самых популярных контролируемых стратегий классификации.

Идея метода довольно проста. Обучающие образцы описываются n-мерными числовыми атрибутами. Учебные образцы хранятся в n-мерном пространстве. Когда дается тестовая выборка с неизвестным классом, классификатор ищет k учебных выборок, которые наиболее близки к неизвестной выборке. При определении близости используются евклидовы расстояния.

Евклидово расстояние находится между двумя точками P (p1, p2,…. Pn) и Q (q1, q2,…. Qn), заданными уравнением 1.

$$d(P,Q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \qquad (1)$$

Для неизвестной выборки данных метка класса назначается путем голосования ее k-ближайших соседей. Несмотря на то, что это довольно простой метод, он имеет некоторые неоспоримые преимущества, такие как:

простота, эффективность, интуитивность, непараметрическая природа и высокая производительность для различных классификационных задач [30]. Однако для больших наборов учебных данных процесс может занимать много времени из-за вычисления расстояния каждого тестового образца до обучающих образцов.

Правильный выбор параметра k, косвенно отвечающего за размер окрестности, и функции расстояния, отвечающей за качество топологического представления, может оказать существенное влияние на результаты [9].

1.2 Наивный байесовский классификатор

Теорема Байеса – основная теорема в теории вероятностей и статистике. Она показывает связь между условными вероятностями и предельными вероятностями для случайных величин [11]. Пусть P (A) будет предшествующей вероятностью, которая является начальной степенью веры в A, а P (B) будет предшествующей вероятностью, которая является начальной степенью веры в B.

P (A | B) - это условная вероятность того, что степень веры в A, принимая во внимание B. Математический вид теоремы Байеса представлен в уравнении 2.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

Наивный байесовский классификатор – это простой вероятностный классификатор, основанный на теореме Байеса в машинном обучении. Пусть D

будет учебным набором базы данных, а X включает в себя n независимых атрибутов (x1, x2, ..., xn), и предположим, что если существует m классов, таких как C1, C2, .., Cm, то классификация состоит в том, чтобы вывести максимальное значение P (Ci | X). Это можно вывести из теоремы Байеса следующим образом [8]. По формуле Байеса (уравнение 3):

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (3)$$

где $P(X|C_i)$ – вероятность встретить объект X среди объектов класса C_i, $P(C_i)$ и $P(X)$ – априорные вероятности класса C_i и объекта X. P(X) не влияет на выбор класса и может быть опущена. Тогда конечная формула для апостериорной вероятности выглядит следующим образом (уравнение 4):

$$P(C_i|X) = P(X|C_i)P(C_i) \qquad (4)$$

Упрощенное предположение в наивном байесовском классификаторе состоит в том, что атрибуты являются равноправными и независимыми. Итак, присвоения классов тестовым образцам основаны на следующих уравнениях:

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \qquad (5)$$

$$\arg\max_{C_i}\left\{P(X|C_i)P(C_i)\right\} \qquad (6)$$

Например, если приходит новая выборка и ее апостериорная вероятность P (C_2 | X) имеет наибольшее значение среди всех P (C_i | X) для всех i классов, она принадлежит классу C_2 в соответствии с теоремой Байеса.

1.3 Алгоритм Случайный лес

1.3.1 Деревья решений как часть Случайного леса

Дерево решений – это контролируемый, непараметрический алгоритм машинного обучения. Его идея состоит в том, чтобы создать набор правил, по которым классифицируемый объект будет отнесен к одному из возможных классов. Дерево решений может объединять как числовые, так и категориальные данные. Пример того, как работает дерево, представлен на рисунке 1:

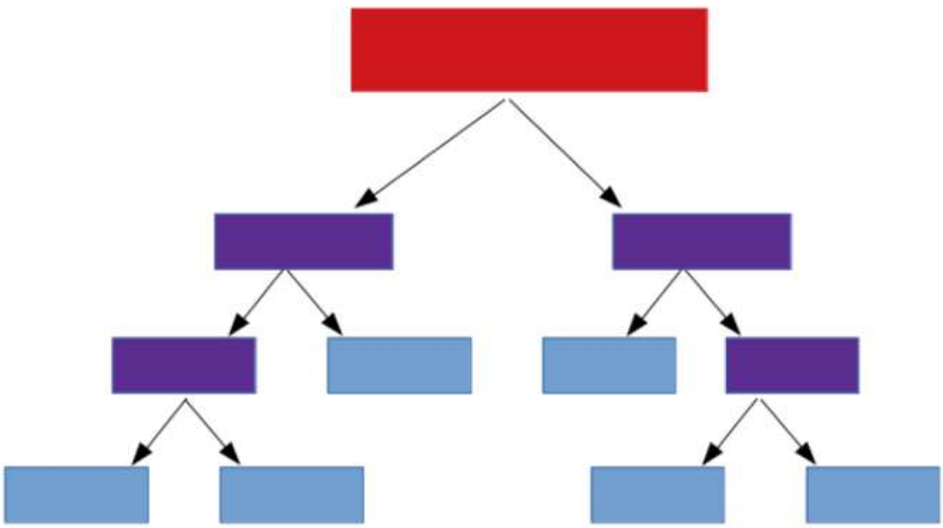


Рисунок 1 – принцип действия дерева решений. Красный цвет – корневой узел, фиолетовый – внутренние узлы, голубой – терминальные узлы

Узлы делятся в зависимости от их положения в дереве решений. Сначала выбирается корневой узел, для этого используется критерий выбора узлов – Индекс Джини, определение которого показано в следующей формуле:

$$I_G = 1 - \sum_{i=1}^{k} p(i|n)^2 \qquad (7)$$

где p(i|n) – доля выборок, которые относятся к классу k для конкретного узла n.

Индекс Джини интерпретируется как критерий ошибочной классификации. При создании структуры дерева решений необходимо выбрать функцию с наименьшим индексом Джини в качестве корневого узла. Так образуются узлы ветвления, для каждого из которых находится такое правило деления, при котором индекс Джини минимальный. Так до тех пор, пока все узлы не станут терминальными узлами. В таких узлах присутствуют представители только одного класса.

Таким образом, классифицируемый объект проходит весь путь от корневого до терминальных узлов и ему назначается класс, который закреплен за тем терминальным узлом, куда попал объект.

1.3.2 Создание алгоритма Случайный лес

Исходя из литературных источников [16] известно, что вероятность корректной классификации ансамблей классификаторов существенно зависит от разнообразия классификаторов, составляющих ансамбль, то есть от того, насколько коррелированны их решения.

В отличие от классических алгоритмов построения деревьев решений [4, 24] в алгоритме Случайный лес при построении каждого дерева на стадиях расщепления вершин используется только фиксированное число случайно отбираемых признаков обучающей выборки [36]. Используя вышесказанные понятия Случайного леса, получаем следующее определение:

Случайный лес – это совокупность множества отдельных деревьев решений, которые могут отличаться друг от друга из-за использования части признаков, отобранных случайно из общего набора признаков, для построения каждого дерева в лесу [23].

База данных, как и для всех алгоритмов, делится на Учебную и Тестовую части (рис. 2). Особенность в том, что Учебная часть делится еще на две для реализации метода «out-of-Bag» (OOB). 2/3 учебных данных используются для построения дерева, а 1/3 для тестирования построенного дерева для оценки его точности.

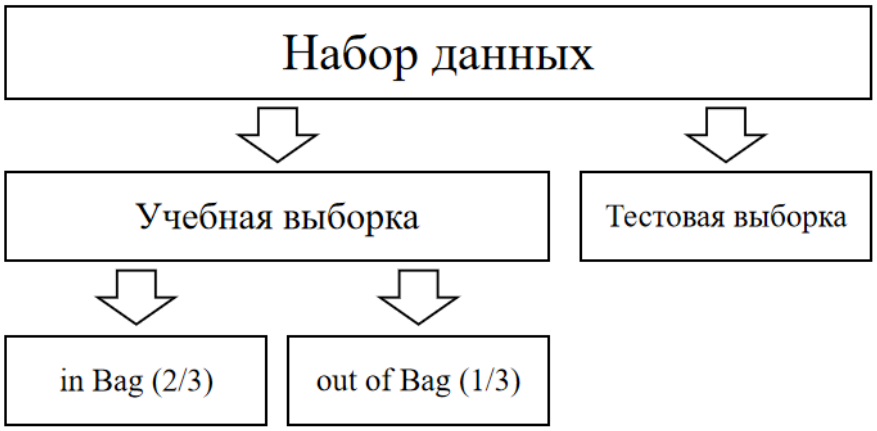


Рисунок 2 – разделение данных при построении алгоритма Случайный лес

Следовательно, легко увидеть, что показатель эффективности в алгоритме Случайного леса зависит от этого метода оценки ошибок.

Классификатор случайных лесов состоит из ансамбля классификаторов дерева решений. Каждое дерево на выходе отдает единичный голос, по его мнению, за верный класс.

Так как эффективность классификации алгоритма случайных лесов зависит от разнообразия деревьев в лесу из-за случайного выбора функций в каждом узле дерева, увеличение признаков использованных в каждом дереве решений, приведет к увеличению, как корреляции, так и силы каждого дерева, что может как улучшить классификацию, так и ухудшить ее. Увеличение же количества деревьев обеспечит более точный классификатор [35]. Одним из эффективных свойств алгоритма случайных лесов является то, что алгоритм не склонен к переобучению, даже если в лес добавлено большое количество деревьев. Брейман утверждает, что деревья всегда сходятся, так что переоснащение не является проблемой [36].

2 Материалы и методы

В данном исследовании реализация перечисленных методов и все манипуляции с данными проводились в свободной среде разработки программного обеспечения с открытым исходным кодом для языка программирования R – RStudio, версия 1.2.5033.

2.1 База данных

Для исследования была использована база данных изучения движения пожилых людей в возрасте от 66 до 86 лет, которые носили пассивную RFID метку для распознавания движений. База данных была предоставлена университетом Аделаиды 12.12.2016 года на Интернет-ресурсе открытого хранилища баз данных для машинного обучения [37] (таблица 1).

Таблица 1 – начальный вид базы данных

64,75	0,60034	0,65224	-0,74348	3	-61	0,24544	921,25	1
65,25	0,62379	0,64076	-0,58384	3	-58	1,7917	922,25	1
65,5	0,62379	0,64076	-0,58384	2	-57,5	5,2662	925,25	1
66	0,40101	0,67521	-0,77769	3	-60,5	0,009204	922,75	1
66,5	0,38928	0,58334	-0,86891	3	-59,5	5,8813	921,75	1
66,75	0,38928	0,58334	-0,86891	3	-60,5	5,5929	925,75	1
67,5	0,7645	0,12402	-0,8233	2	-59	1,557	923,75	3
69,5	1,1163	0,16995	-0,447	3	-58,5	5,9457	920,75	3
69,75	1,1397	0,1355	-0,35577	3	-59,5	4,0727	920,25	3
70	1,1397	0,1355	-0,35577	3	-59,5	1,0861	923,75	3
70,75	1,1397	0,1355	-0,35577	3	-60,5	0,76392	923,75	3

Значения данных:

Столбец 1: Время в секундах

Столбец 2: Показание ускорения в G для передней оси

Столбец 3: Показание ускорения в G для вертикальной оси

Столбец 4: Показание ускорения в G для боковой оси

Столбец 5: Идентификатор датчика считывания антенны

Столбец 6: Уровень принимаемого сигнала

Столбец 7: Фаза

Столбец 8: Частота

Столбец 9: Состояния испытуемых:

1 – Сидит на кровати;

2 – Сидит на стуле;

3 – Лежит;

4 – Передвигается.

2.2 Первичная обработка данных

База данных представляет собой 87 файлов расширения csv, из них 24 файла содержат данные эксперимента испытуемых мужского пола, а 63 женского. Для исследования все файлы были объединены в один. Столбец «время» был заменен на столбец «пол», где «0» это женский пол, а «1» мужской.

Строки с частично отсутствующими и аномальными значениями были удалены из базы данных.

Данные были поделены на две выборки:

1) Обучающая (63122 векторов);

2) Проверочная (11237 векторов).

В свою очередь, обучающая выборка была поделена на 10 равных обучающих подвыборок (ОП) размером по 6312 векторов, а проверочная выборка на 18 проверочных подвыборок для будущих оценок устойчивости алгоритмов.

2.3 Анализ данных

Была составлена корреляционная матрица (табл. 2). По данным таблицы был сделан вывод о весомой корреляции (>|0.3|) трех предикторов:

1) Показание ускорения в G для передней оси;

2) Показание ускорения в G для вертикальной оси;

3) Показание ускорения в G для боковой оси.

Таблица 2 – корреляционная матрица

	1	2	3	4	5	6	7	8	9
1	1	0,202	-0,016	0,435	-0,106	-0,102	0,054	0,014	-0,147
2	0,202	1	-0,731	0,378	0,134	-0,114	0,069	0,064	0,415
3	-0,016	-0,731	1	0,162	-0,261	0,004	-0,028	-0,020	-0,75
4	0,435	0,378	0,162	1	-0,047	-0,186	0,078	0,113	-0,335
5	-0,106	0,134	-0,261	-0,047	1	-0,185	0,031	0,048	0,204
6	-0,102	-0,114	0,004	-0,186	-0,185	1	0,004	-0,211	0,124
7	0,054	0,069	-0,028	0,078	0,031	0,004	1	-0,050	-0,006
8	0,014	0,064	-0,020	0,113	0,048	-0,211	-0,050	1	-0,005
9	-0,147	0,415	-0,750	-0,335	0,204	0,124	-0,006	-0,005	1

Для анализа данных этих предикторов база данных была поделена на четыре части по состояниям испытуемых, то есть первая часть содержит данные только первого состояния испытуемых, вторая часть только второго состоянии и т.д. После чего были построены гистограммы каждого из этих

предикторов в каждом состоянии (Приложение А). Параметры гистограмм показаны в таблице 3.

Таблица 3 – Параметры гистограмм

Состояние	Ускорение	Среднее	Медиана	Дисперсия	Ст. отклонение	Мин. значение	Макс. значение	Коэф. ассиметрии
1	1	0,351	0,342	0,04	0,201	-0,56	1,234	0,185
	2	0,934	0,951	0,011	0,106	-0,152	2,03	-2,923
	3	0,033	0,032	0,023	0,151	-1,12	0,9	-1,911
2	1	0,54	0,589	0,066	0,256	-0,666	1,128	-0,79
	2	0,868	0,87	0,01	0,099	0,526	1,33	0,311
	3	0,068	0,089	0,013	0,111	-0,344	0,488	-0,701
3	1	0,836	1,023	0,154	0,393	-0,244	1,468	-0,499
	2	0,099	0,055	0,031	0,176	-0,553	1,066	0,208
	3	-0,362	-0,116	0,24	0,49	-1,336	1,218	-0,163
4	1	0,227	0,237	0,05	0,223	-0,748	1,105	-0,092
	2	0,959	0,974	0,015	0,123	-0,278	1,548	-1,133
	3	0,014	0,009	0,023	0,152	-0,778	0,91	0,632

2.4 Классификаторы

Результатом классификаторов является столбец предсказанных состояний. Для определения точности было сделано сравнение результатов классификаторов и действительных значений состояний. Если состояния совпадали, классификация считалось верной. Точность определялась как доля верной классификации.

2.4.1 Алгоритм k-ближайших соседей

Для классификации методом k-ближайших соседей была использована библиотека Class, функция knn. Первым аргументом в функцию была передана обучающая выборка. Второй аргумент – проверочная выборка без столбца класса. Третьим аргументом в функцию передается, отдельно от проверочной выборки, вектор классов обучающей выборки. Четвертый аргумент – это параметр k, который отвечает за количество соседей, чьи голоса необходимо учитывать в присвоении класса классифицируемому объекту.

2.4.2 Наивный байесовский классификатор

Для классификации с использованием апостериорных вероятностей был применен Наивный байесовский классификатор, допуская, что все предикторы являются равноправными. Для этого была использована библиотека Naivebayes, функции naive_bayes и predict. В функцию naive_bayes первым аргументом передается обучающая выборка без столбца классов. Вторым аргументом был передан вектор классов обучающей выборки. Сглаживание Лаплас (laplace = 1) необходимо для классификации объекта, чьи значения предикторов неизвестны

классификатору. Ядерное сглаживание (usekernel = T) использовалось для оценки плотности предикторов. Распределение Пуассона (usepoisson = T) необходимо для модели наивного байесовского классификатора, в которой все условные распределения класса предполагаются пуассоновскими и независимыми. В функцию predict первым аргументом передавался результат функции naive_bayes. Вторым аргументом – проверочная выборка.

2.4.3 Алгоритм Случайный лес

Для классификации алгоритмом Случайный лес была использована библиотека RandomForest, функции randomforest и predict. В функцию randomforest первым аргументом передается обучающая выборка без столбца классов. Вторым аргументом необходимо передать вектор значений классов обучающей выборки. Остальные аргументы требовали присвоения значений. Такими важными аргументами являются:

ntree – количество решающих деревьев в лесу;

mtry – количество случайно выбранных предикторов при каждом разделении решающего дерева;

nodesize – минимальный размер терминальных узлов. Увеличение этого числа приводит к выращиванию деревьев меньшего размера;

importance – учитывание влияния предикторов (необходимый аргумент в реальных задачах, так как предикторы имеют разное влияние для присвоения класса классифицируемому объекту);

keep.forest – необходим для сохранения леса для использования в дальнейших классификациях.

В функцию predict передается первым аргументом результат функции randomforest. Вторым аргументом передается проверочная выборка. Значение аргумента predict.all = TRUE необходимо для вывода голосов всех деревьев в лесу.

3 Результаты и обсуждения

3.1 Алгоритм k-ближайших соседей

Для определения оптимального значения параметра k (количество ближайших векторов обучающей выборки, которые участвуют в присвоении класса неизвестному вектору) был построен график зависимости точности от величины параметра k (рис.3). По этой зависимости можно сделать вывод, что увеличение k приводит к уменьшению точности.

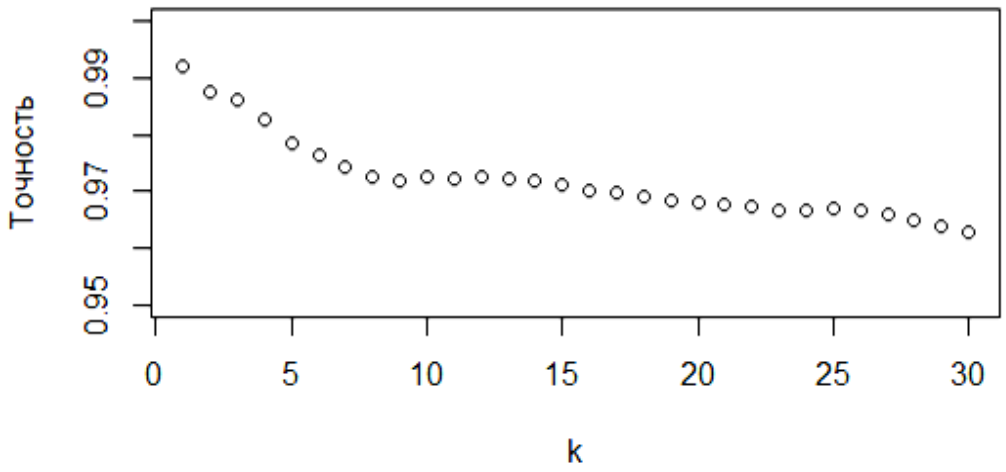


Рисунок 3 – зависимость точности классификации от величины параметра k

Для изучения производительности метода классификации определили закономерность увеличения времени классификации от величины параметра k (рис. 4). Из которой видно, что увеличение параметра k влечет за собой увеличение времени классификации.

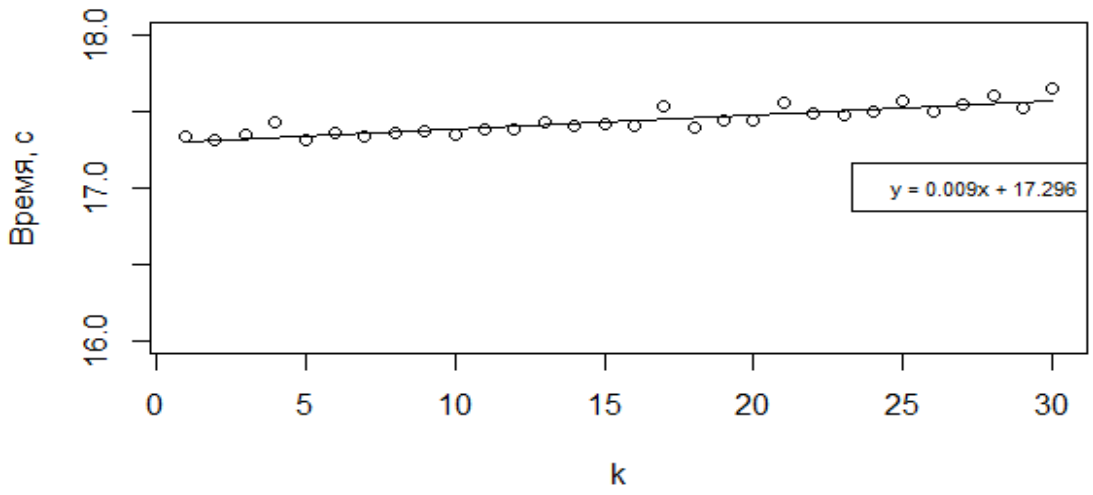


Рисунок 4 – зависимость времени классификации от величины параметра k

Была определена зависимость времени классификации от размера обучающей выборки (рис. 5). По этой зависимости делаем вывод, что алгоритм имеет увеличение времени классификации на 0.27 секунд за каждую тысячу значений обучающей выборки, что требует найти оптимальный размер обучающей выборки.

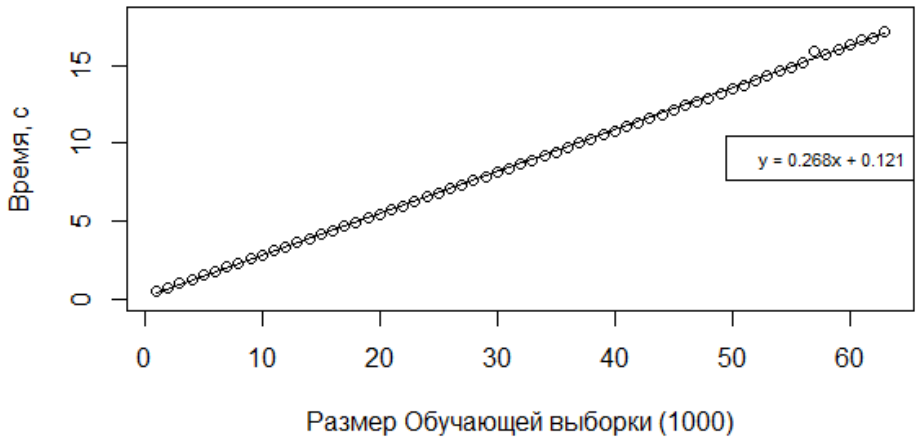


Рисунок 5 – зависимость времени классификации проверочной выборки от размера обучающей выборки

Для оптимизации метода путем выбора оптимальной величины обучающей выборки был построен график зависимости точности всей проверочной выборки от размера обучающей выборки (рис. 6). При размере обучающей выборки 30 тысяч значений точность выходит на плато. То есть использование обучающей выборки более 30 тысяч значений нецелесообразно, так как мы имеем увеличение затраченного времени на классификацию без улучшения точности.

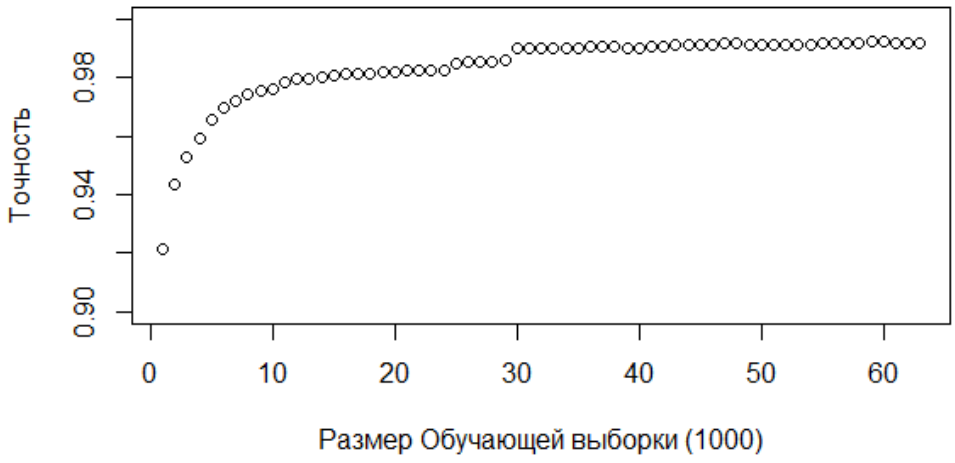


Рисунок 6 – зависимость точности классификации проверочной выборки от размера обучающей выборки

Был построен точечный график, где точки показывают точность классификации всех комбинаций ОП и проверочных подвыборок (рис. 7). Это было необходимо сделать для оценки устойчивости алгоритма. Из этого графика видно, что семнадцать из восемнадцати проверочных подвыборок сходятся при всех ОП. Двенадцатая проверочная подвыборка имеет небольшое расхождение, некритичное для таких небольших размеров ОП.

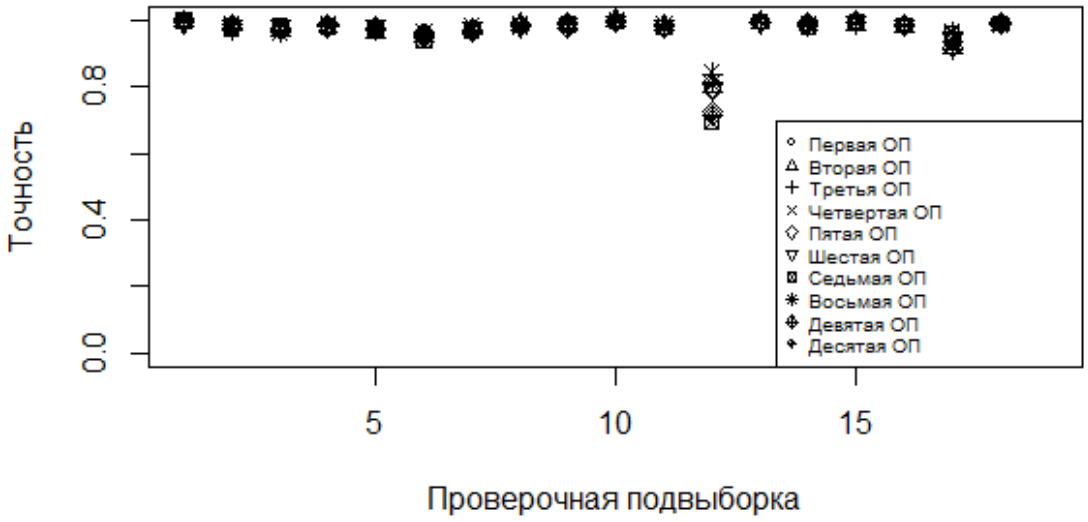


Рисунок 7 – точечный график классификаций всех комбинаций ОП и проверочных подвыборок

3.2 Наивный байесовский классификатор

Первичными априорными вероятностями были использованы пропорции классов обучающей выборки:

1 - 0.23;

2 - 0.06;

3 - 0.68;

4 - 0.03 .

Была найдена лучшая комбинация аргументов, требующих присвоения значений, путем построения точечного графика (рис. 8). На графике точками отображена точность классификаций восемнадцати проверочных подвыборок в

разных комбинациях использования/неиспользования ядерного сглаживания и использования/неиспользования распределения Пуассона. Выделяются 2 подвыборки (2 и 7) на которых отчетливо видно, что использование распределения Пуассона приводит к существенному падению точности. На большинстве остальных подвыборок хоть и в меньшей степени, но тоже хуже, чем в комбинациях, где распределение Пуассона не используется. Ядерное сглаживание наоборот, приводит к улучшению или неизменению точности, что говорит о целесообразности его использования.

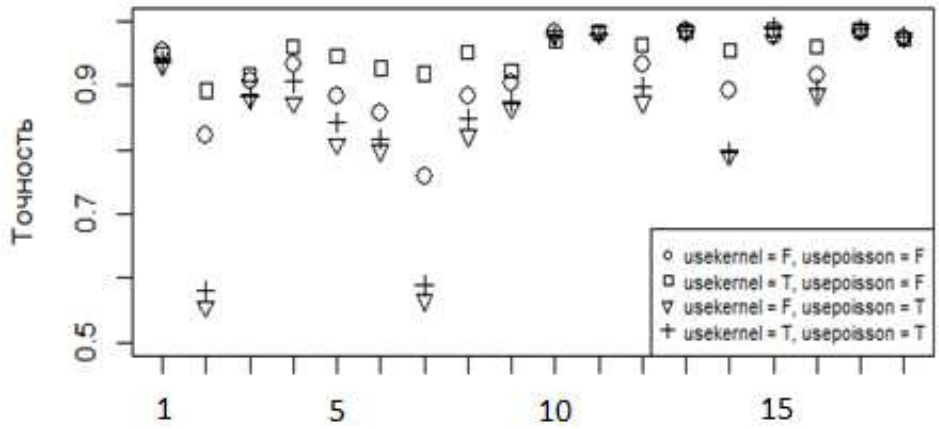


Рисунок 8 – точность классификаций восемнадцати проверочных подвыборок в разных комбинациях использования/неиспользования ядерного сглаживания и использования/неиспользования распределения Пуассона

По рисунку 9 можно сделать вывод о неразличимости затраченного времени на классификацию проверочной выборки в комбинациях аргументов перечисленных выше.

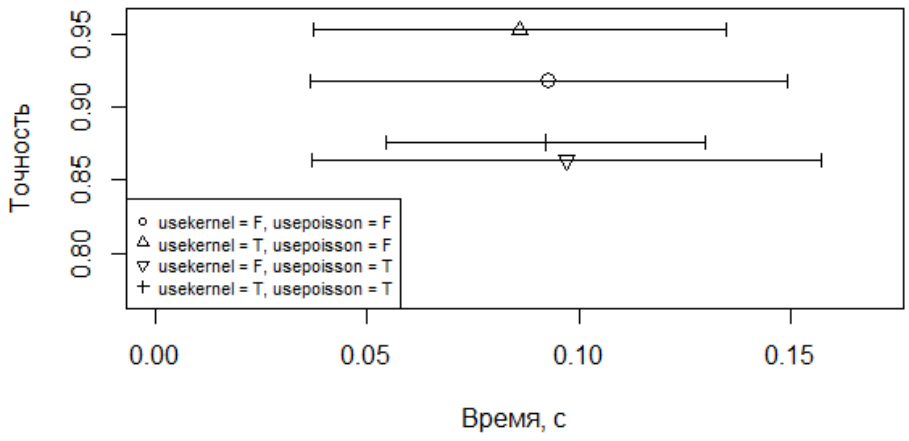


Рисунок 9 – Время выполнения классификации в разных комбинациях аргументов usekernel и usepoisson

Был построен точечный график, где точки показывают точность классификации всех комбинаций ОП и проверочных подвыборок (рис. 10). Это было необходимо сделать для оценки устойчивости алгоритма.

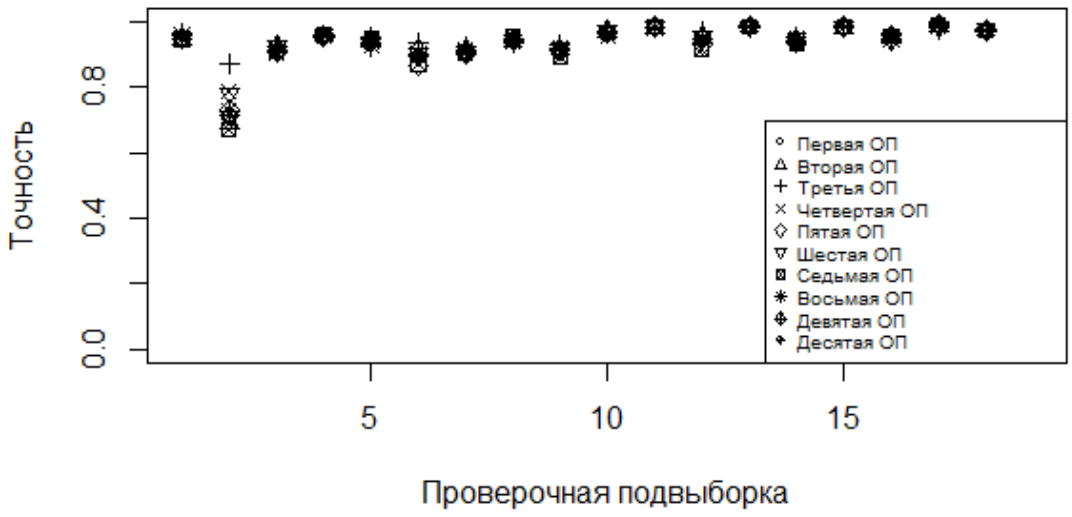


Рисунок 10 – точечный график классификаций всех комбинаций ОП и проверочных подвыборок

Из этого графика видно, что семнадцать из восемнадцати проверочных подвыборок сходятся при всех ОП. Вторая проверочная подвыборка имеет небольшое расхождение.

3.3 Алгоритм Случайный лес

Было проведено варьирование двух основных аргументов с целью подобрать лучшие для данной задачи классификации (рис. 11). На вертикальной оси расположена точность, а на горизонтальной параметр размера леса. Для того чтобы сделать основные выводы по данному графику, достаточно четырех значений параметра размера дерева, которые указаны на графике. Мы делаем три важных вывода:

1) Независимость параметров друг от друга;

2) Быстрый выход на плато по параметру Forestsize (50+ деревьев);

3) Существенная зависимость от параметра k.

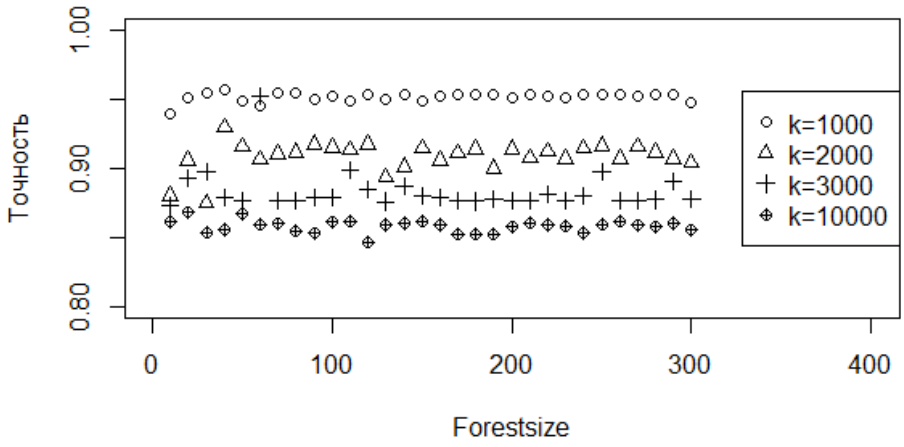


Рисунок 11 – зависимость точности от параметров k и Forestsize

Для поиска зависимости от размера дерева и нахождения лучшего размера был построен график (рис. 12), на котором по вертикальной оси для черных точек находится точность классификации, а для зеленых треугольников указана сумма квадратов отклонений (СКО) поднятая на 0.94. На горизонтальной оси находится параметр k, который влияет на размер отдельного дерева в лесу.

По точкам на графике можно заметить тот эффект переподгонки. При k<83 наблюдается улучшение точности обученных моделей, а на участке k> 83 ухудшение. Для нашей задачи лучше всего подходят обрезанные деревья. При k=83 СКО достигает минимума, а точность максимума, из-за чего в итоговой классификации используется это значение аргумента nodesize.

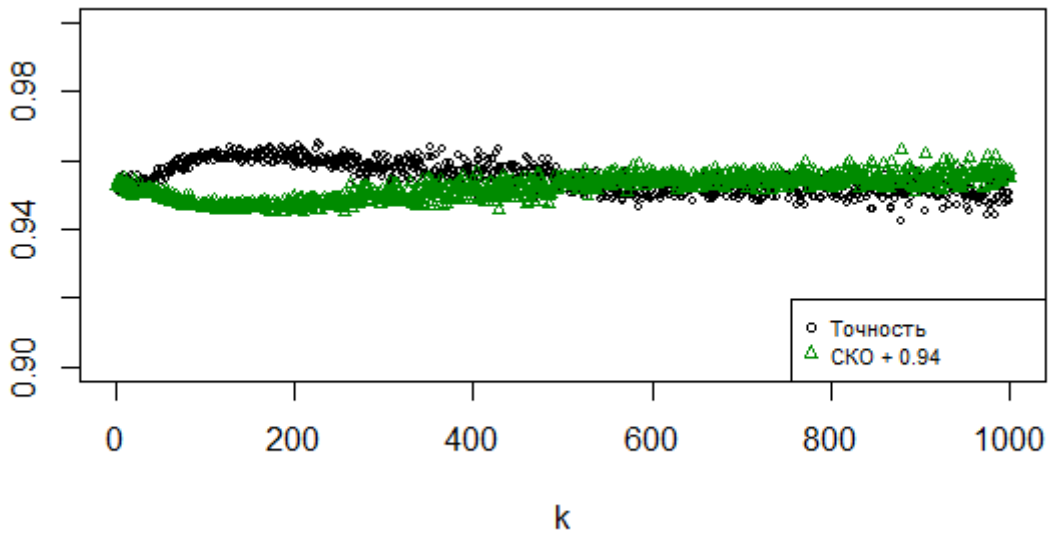


Рисунок 12 – зависимость точности и СКО от параметра k

Была определена зависимость времени классификации от размера леса. Из графика (рис. 13) видно, что при увеличении леса на одно дерево повлечет за собой увеличение времени классификации на 0.08 секунд. Вследствие чего,

«выращивание» леса с более чем 50 деревьев нецелесообразно. В связи с этим, в итоговой классификации использовался аргумент Forestsize равный 50.

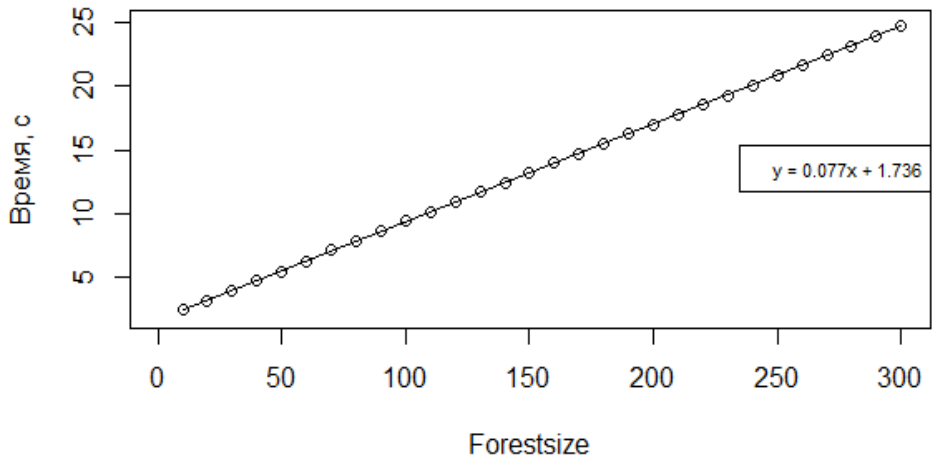


Рисунок 13 – зависимость времени классификации от размера леса

Был построен точечный график, где точки показывают точность классификации всех комбинаций ОП и проверочных подвыборок (рис. 14). Это было необходимо сделать для оценки устойчивости алгоритма. Из этого графика видно, что алгоритм имеет устойчивость хуже, чем у других алгоритмов данного исследования, при таком размере ОП.

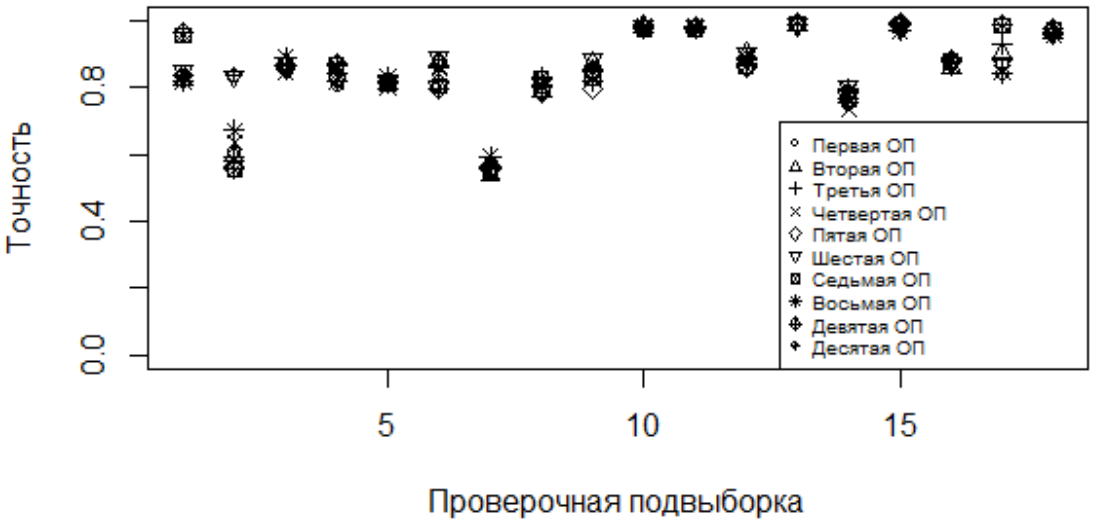


Рисунок 14 – точечный график классификаций всех комбинаций ОП и проверочных подвыборок

3.4 Сравнение алгоритмов классификации

На результирующий график (рис. 15) были вынесены точности классификаций проверочных подвыборок тремя алгоритмами классификации, описанные выше. В каждом алгоритме использовались лучшие аргументы, которые были определены в данном исследовании.

Метод k-ближайших соседей показал лучший результат значений точности и устойчивости.

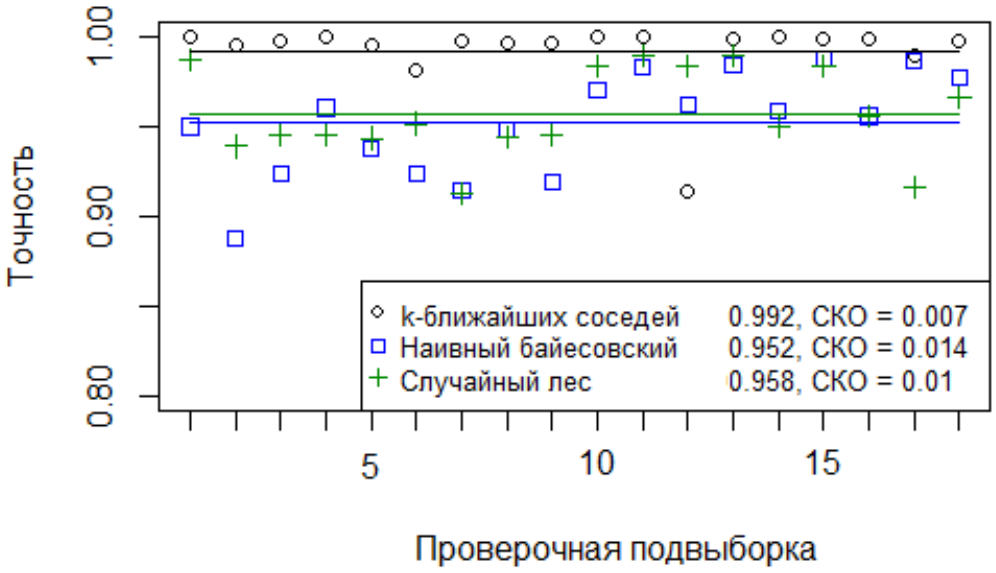


Рисунок 15 – точность классификации проверочных подвыборок тремя алгоритмами классификации

ЗАКЛЮЧЕНИЕ

1. Из литературных источников выделены методы классификации, подходящие к данной задаче;

2. Решена задача классификации тремя разными алгоритмами (k-ближайших соседей, Наивный байесовский классификатор, Случайный лес);

3. Проведен анализ начальных данных и найдены лучшие параметры алгоритмов;

4. Проведено сравнение алгоритмов между собой: алгоритм k-ближайших соседей является лучшим для задачи классификации определения положения изолированных пациентов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Bao, L.; Intille, S. S. Activity recognition from user-annotated acceleration data / L. Bao, S. S. Intille // in Pervasive Computing, ser. LNCS. Springer. – 2004. – №3001. – P. 1 – 17.

2. Bianchi, F.; Redmond, S. J.; Narayanan, M. R.; Cerutti, S.; Lovell, N. Barometric pressure and triaxial accelerometry-based falls event detection / F. Bianchi, S. J. Redmond, M. R. Narayanan, S. Cerutti, N. Lovell // IEEE Trans. Neural Syst. Rehabil. Eng. – 2010. – vol. 18. – №6. – P. 619 – 627.

3. Breiman, L. Random forests / L. Breiman // Machine Learning 45. – 2001. – P. 5 – 32.

4. Breiman L.; Friedman R.; Olshen R.; Stone C. Classification and Regression Trees. Belmont / L. Breiman, R. Friedman, R. Olshen, C. Stone // California: Wadsworth International. – 1984. – 342 p.

5. Bruyneel, M.; Libert, W.; Ninane, V. Detection of bed-exit events using a new wireless bed monitoring assistance / M. Bruyneel, W. Libert, V. Ninane // Int. J. Med. Inform. – 2011. – vol. 80. – №2. – P. 127 – 132.

6. Capezuti, E.; Brush, B.; Lane, S.; Rabinowitz, H.; Secic, M. Bed-exit alarm effectiveness / E. Capezuti, B. Brush, S. Lane, H. Rabinowitz, M. Secic // Arch. Gerontol. Geriatr. – 2009. – vol. 49. – №1. – P. 27 – 31.

7. Doukas, C.; Maglogiannis, I. Advanced patient or elder fall detection based on movement and sound data / C. Doukas, I. Maglogiannis // in Proceedings of the Second International Conference on Pervasive Computing Technologies for Healthcare. – 2008. – P. 103 – 107.

8. Duda, R. O.; Hart, P. E. Pattern Classification and Scene Analisis / R. O. Duda, P. E. Hart // Wiley, New York. – 1973. – vol. 3.

9. Gou, J.; Du, L.; Zhang, Y.; Xiaong, T. A new distance-weighted k-nearest neighbor classifier / J. Gou, L. Du, Y. Zhang // J. Inf. Comput. Sci.– 2012. –№9(6). – P. 1429 – 1436.

10. Govercin, M.; Moltzsch, Y. K.; Meis, M.; Wegel, S.; Gietzelt, M.; Spehr, J.; Winkelbach, S.; Marschollek, M.; Steinhagen-Thiessen, E. Definingthe user requirements for wearable and optical fall prediction and falldetection devices for home use / M. Govercin, Y. K. Molzsch, M. Meis, S. Wegel, M. Gietzelt, J. Spehr, S. Winkelbach, M. Marschollek, E. Steinhagen-Thiessen // Informatics for Health and Social Care. – 2010. – vol. 35. – №3-4. – P. 177 – 187.

11. Hanson, R.; Stutz, J.; Cheeseman, P. Bayesian Classification Theory / R. Hanson, J. Stutz, P. Cheeseman // NASA Ames Research Center, Artificial Intelligence Branch. – 1991.

12. Hilbe, J.; Schulc, E.; Linder, B.; Them, C. Development and alarm threshold evaluation of a side rail integrated sensor technology for the prevention of falls / J. Hilbe, E. Schulc, B. Linder // Int. J. Med. Inform. – 2010. – vol. 79. – №3. – P. 173 – 180.

13. Hill, K.; Vu, M.; Walsh, W. Falls in the acute hospital setting impact on resource utilisation / K. Hill, M. Vu, W. Walsh // Aust. Health Rev. – 2007. – vol. 31. – №3. – P. 471 – 477.

14. Hitcho, B.; Krauss, M. J.; Birge, S.; Claiborne Dunagan,W.; Fischer, I.; Johnson, S.; Nast, P. A.; Costantinou, E.; Fraser, V. J. Characteristics and circumstances of falls in a hospital setting / B. Hitcho, M. J. Krauss, S. Birge, W. Claiborne Dunagan, I. Fischer, S / Johnson, P. A. Nast // J. Gen. Intern. Med. – 2004. – №19. – P. 732 – 739.

15. Kaufmann, T.; Ranasinghe, D. C.; Zhou, M.; Fumeaux, C. Wearable quarter-wave folded microstrip antenna for passive UHF RFID applications / T. Kauffman, D. C. Ranasinghe, M. Zhou, C. Fumeaux // Int. J. Antennas Propag. – 2013.

16. Kuncheva L. I. Combining Pattern Classifiers: Methods and Algorithms / L. I. Kuncheva // Hoboken, New Jersey: John Wiley Sons. – 2004. – 349 p.

17. Lai, C. F.; Chang, S. Y.; Chao, H. C.; Huang, Y. M. Detection of cognitive injured body region using multiple triaxial accelerometers for elderly falling / C. F. Lai, S. Y. Chang, H. C. Chao, Y. M. Huang // IEEE Sensors J.– vol. 11. – №3. – P. 763 – 770.

18. Mubashir, M.; Shao, L.; Seed, L. A survey on fall detection: Principles and approaches / M. Mubashir, L. Shao, L. Seed // Neurocomputing. – 2013. – vol. 100. – P. 144– 152.

19. Mukhopadhyay, S. Wearable sensors for human activity monitoring: Areview / S. Mukhopadhyay // IEEE Sensors Journal. – 2015. – vol. 15. –№3. P. 1321 – 1330.

20. Oliver, D.; Prevention of falls in hospital inpatients. agendas for research and practice / D. Oliver // Age Ageing. – 2004. – vol. 33. – №4. – P. 328 – 330.

21. Oliver, D.; Hopper, A.; Seed, P. Do hospital fall prevention programs work? A systematic review / D. Oliver, A. Hopper, P. Seed // J. Am. Geriatr. Soc. – 2000. – №48. – P. 1679 – 1689.

22. Ojetola, O.; Gaura, E.; Brusey, J. Fall detection with wearable sensors–safe (smart fall detection) / O. Ojetola, E. Gaura, J. Brusey // in 2011 7th Int. Conf. on Intelligent Environments. – 2011. – P. 318 – 321.

23. Pater, N.; Enchanting Random Forest Implementation in Weka / N. Pater // Machine Learning Conference Paper for ECE591Q. – 2005.

24. Quinlan J. R. Simplifying decision trees / J. R. Quinlan // InternationalJournal of Man Machine Studies. – 1987. –Vol. 27. – P. 221 – 234.

25. Sample, A. P.; Yeager, D. J.; Powledge, P. S.; Mamishev, A. V.; Smith, J. R. Design of an RFID-based battery-free programmable sensingplatform / A. P.

Sample, D. J. Yeager, P. S. Powledge, A. V. Mamishev, J. R. Design // IEEE Trans. Instrum. Meas. – 2008. – vol. 57. – №11. – P. 2608 – 2615.

26. Schwarz, L. A.; Mateus, D.; Navab, N. Recognizing multiple human activities and tracking full-body pose in unconstrained environments / L. A. Schwarz, D. Mateus, N. Navab // Pattern Recognition. – 2012. – vol. 45. –№1. – P. 11 – 23.

27. Shany, T.; Redmond, S.; Narayanan, M.; Lovell, N. Sensors-based wearable systems for monitoring of human movement and falls / T. Shany, S. Redmond, M. Narayanan, N. Lovell // IEEESensors J. – 2012. – vol. 12. – №3. – P. 658 – 670.

28. Stikic, M.; Larlus, D.; Ebert, S.; Schiele, B. Weakly supervised recognition of daily life activities with wearable sensors / M. Stikic, D. Larlus, S. Ebert, B. Schiele // IEEE Trans.Pattern Anal. Mach. Intell. – 2011. – vol. 33. – №12. – P. 2521 – 2537.

29. Tideiksaar, R.; Feiner, C. F.; Maby, J. Falls prevention: the efficacy of a bed alarm system in an acute-care setting / R. Tideiksaar, C. F. Feiner, J. Maby // Mt. Sinai J. Med. – 1993. –vol. 60. – №6. – P. 522 – 527.

30. Torralba, A.; Fergus, R.; Freeman, W. T. 80 million tiny images: a large data set for nonparametric object and scene recognition / A. Torralba, R. Fergus, W. T. Freeman // PAMI. – 2008. – №30(11). – 1958 – 1970.

31. Vass, C. D.; Sahota, O.; Drummond, A.; Kendrick, D.; Gladman, J.; Sach, T.; Avis, M.; Grainge, M. REFINE (Reducing falls in inpatient elderly)-a randomised controlled trial / C. D. Vass, O. Sahota, A. Drummond, D. Kendrick, J. Gladman, T. Sach, M. Avis, M. Grainge // Trials. – 2009. – vol. 10. – №1. – P. 83.

32. Vassallo, M.; Amersey, R.A.; Sharma, J.C.; Allen, S.C. Falls on integrated medical wards / M. Vassallo // Gerontology. – 2000. – №46. – P. 158 – 162.

33. Wolf, K. H.; Hetzer, K.; Schwabedissen, H. M.; Wiese, B.; Marschollek, M. Development and pilot study of a bed-exit alarmbased on a body-worn

accelerometer / K. H. Wolf, K. Hetzer, H. M. Schwabedissen, B. Wiese, M. Marschollek // Zeitschrift fur Gerontologie und Geriatrie. – 2013. – vol. 46. – №8. – P. 727 – 733.

34. Wolpert, D. H. The lack of a priori distinctions between learning algorithms / D. H. Wolpert // Neural Computation. – 1996. – vol. 8. – №7. – P. 1341 – 1390.

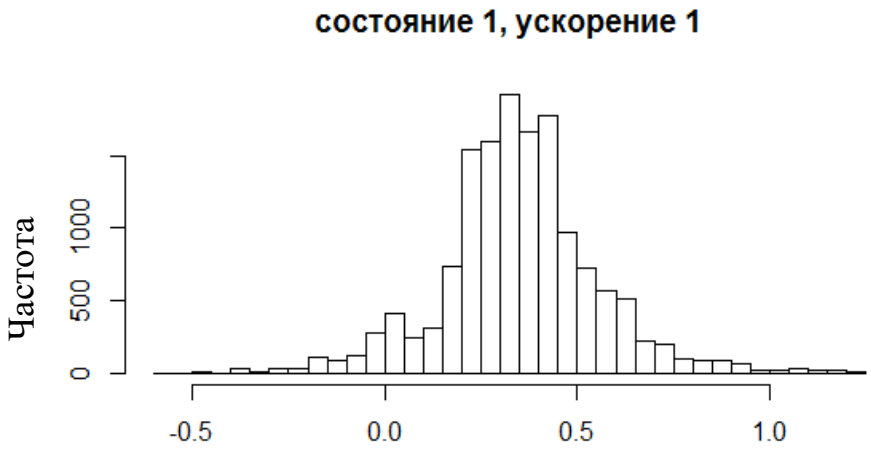
35. Yuan, J.; Tan, K.; Lee, T.; Koh, G. Power-efficient interrupt-driven algorithms for fall detection and classification of activities of dailyliving / J. Yuan, K. Tan, T. Lee, G. Koh // IEEE Sensors J.– 2015. –vol. 15. – №3. – P. 1377 – 1387.

36. Труды Карельского научного центра РАН № 1. – 2013.– С. 117 – 136 УДК 512.2 СЛУЧАЙНЫЕ ЛЕСА: ОБЗОР С. П. Чистяков.

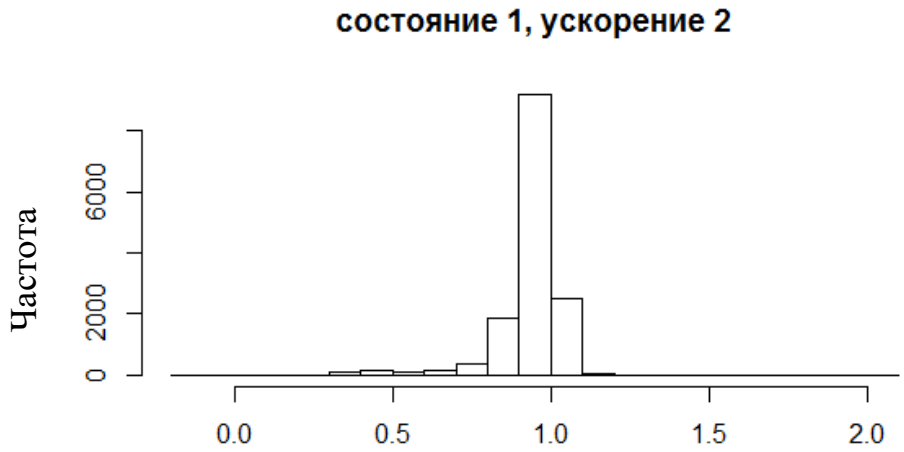
37. Machine Learning Repository [Электронный ресурс]: Activity recognition with healthy older people using a batteryless wearable sensor Data Set / 12.12.2016 // Center for Machine Learning and Intelligent Systems. – https://archive.ics.uci.edu/ml/datasets/Activity+recognition+with+healthy+older+people+using+a+batteryless+wearable+sensor.

ПРИЛОЖЕНИЕ А

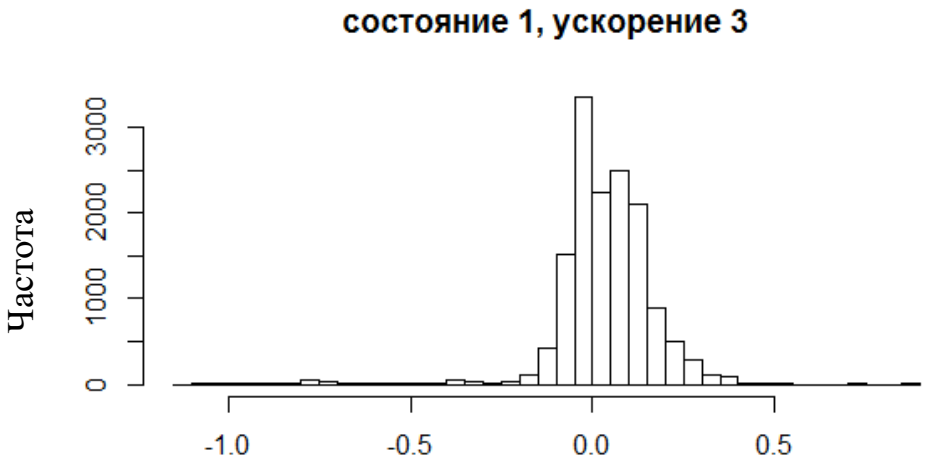
Гистограммы выбранных предикторов в каждом состоянии



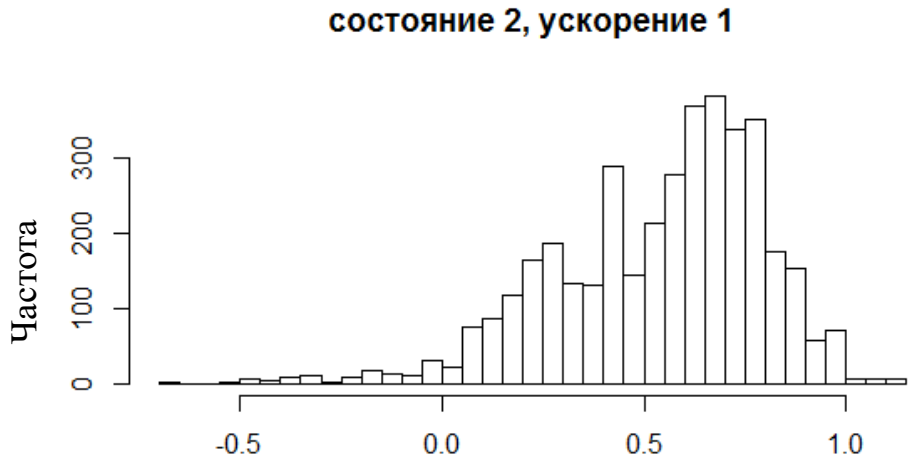
Значения ускорений в G



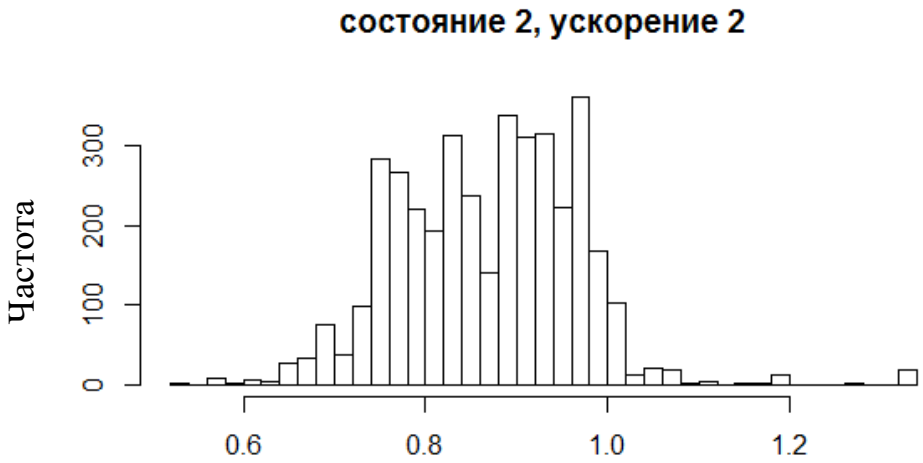
Значения ускорений в G



Значения ускорений в G

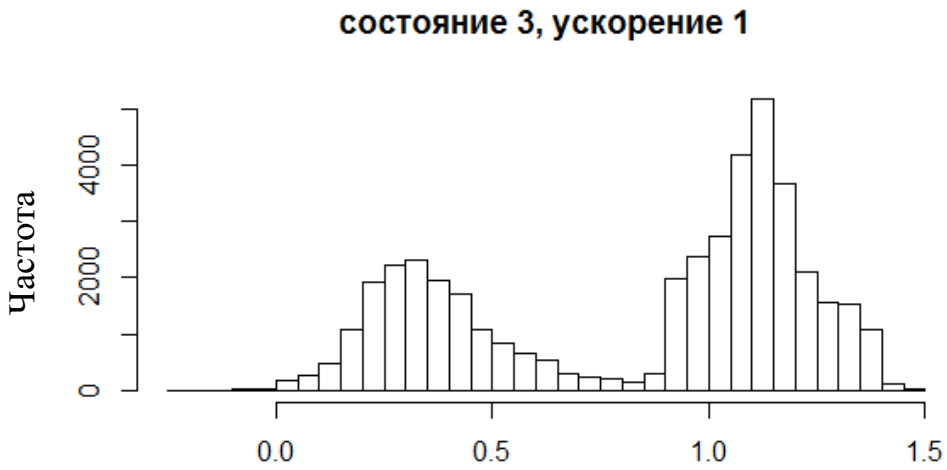


Значения ускорений в G



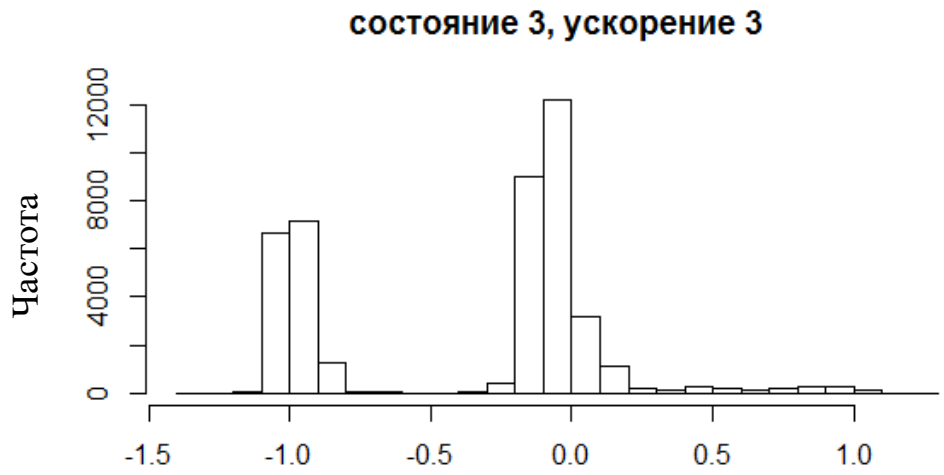
Значения ускорений в G

Значения ускорений в G

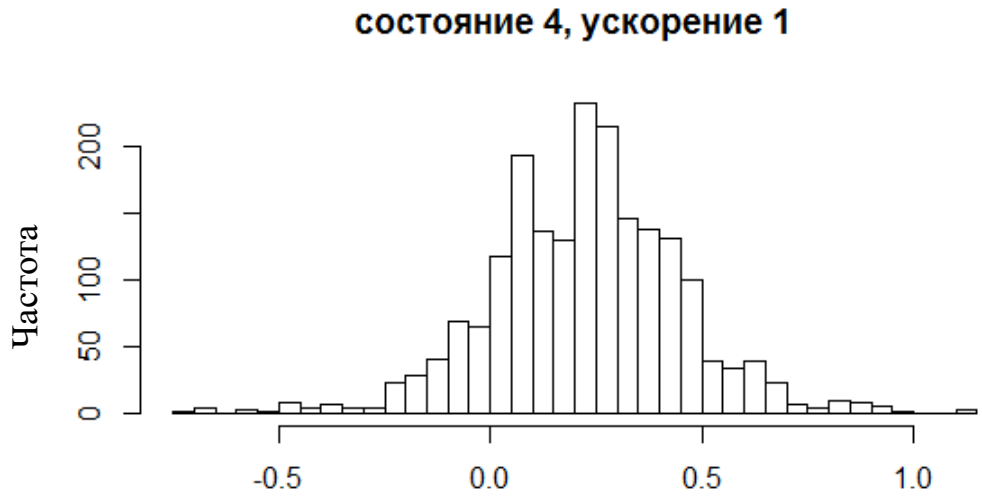


Значения ускорений в G

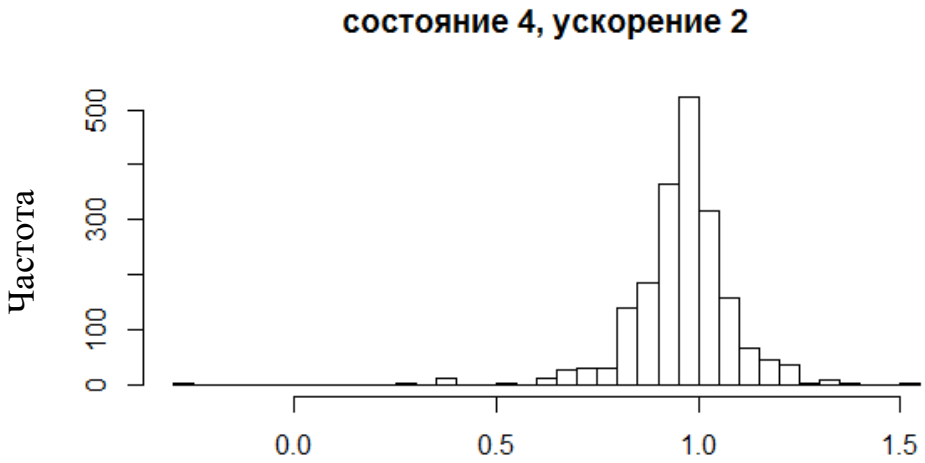
Значения ускорений в G



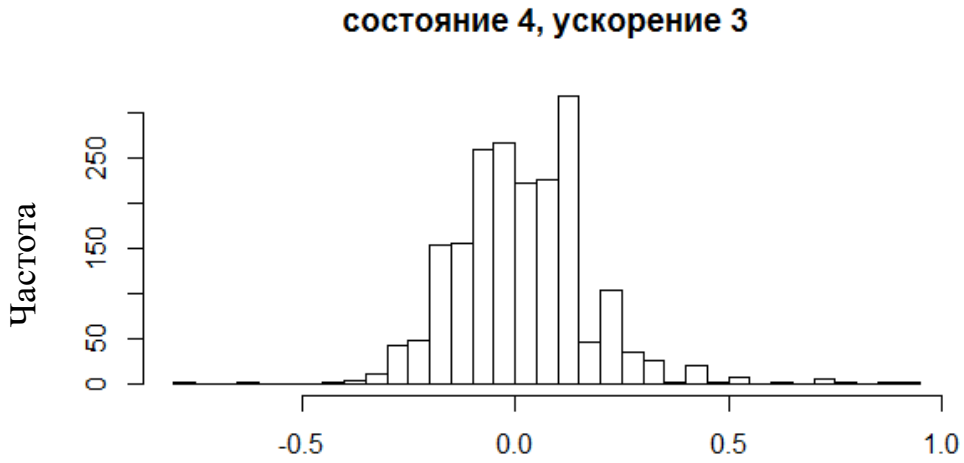
Значения ускорений в G



Значения ускорений в G



Значения ускорений в G



Значения ускорений в G

Федеральное государственное автономное
образовательное учреждение высшего
образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии

Базовая кафедра медико-биологических систем и комплексов

УТВЕРЖДАЮ
Заведующий кафедрой
_____ А.Н. Шуваев
« 22» июня 2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Сравнения алгоритмов классификации для определения положения изолированных пациентов

03.04.02 Физика
03.04.02.09 Технологическое сопровождение ядерной медицины и медицинского оборудования

Научный руководитель _____ 22.06.2020г. к.ф-м.н. А.Н. Шуваев

Выпускник _____ 22.06.2020г. А.Ю. Чмурин

Рецензент _____ 22.06.2020г. к.б.н., доцент Л.В. Степанова

Красноярск 2020