Санкт-Петербургский государственный университет

Информационно-аналитические системы

Группа 22.М04-мм

Екатерина Петровна Винник

Прогнозирование отказов жестких дисков системы Tatlin.Unified

Отчет по производственной практике

Научный руководитель:
к.ф.-м.н., доц. Е.Г. Михайлова

Консультант:
к.ф.-м.н., В.И. Гориховский

Санкт-Петербург
2023

Saint Petersburg State University

Software and Administration of Information Systems

Group 22.M04-mm

Ekaterina Vinnik

Failure Prediction in Tatlin.Unified Hard Disk Drives

Internship report

Scientific supervisor:
C.Sc, docent. E.G. Mikhailova

Consultant:
C.Sc, V. I. Gorikhovskii

Saint Petersburg
2023

Оглавление

Введение **4**

Постановка задачи **6**

1. Существующие подходы к прогнозированию отказа жесткого диска **7**

1.1. Прогнозирование отказа жесткого диска с помощью решения задачи бинарной классификации 7

1.2. Прогнозирование отказа жесткого диска с помощью решения задачи выявления аномалий 9

1.3. Прогнозирование отказа жесткого диска с помощью решения задачи кластеризации 12

2. Применение подходов на наборе модельных данных **17**

2.1. Выбор набора модельных данных 17

2.2. Применение алгоритмов бинарной классификации 18

2.3. Применение алгоритмов выявления аномалий 21

2.4. Применение алгоритмов кластеризации 22

2.5. Анализ примененных подходов 26

Заключение **27**

Список литературы **29**

Введение

Жесткие диски являются одним из самых распространенных устройств хранения и присутствуют в разных системах — от персональных компьютеров до систем хранения данных. Согласно [9], большинство информации, производимой в мире, хранится на жестких дисках. Частота отказов жесткого диска в год составляет чуть меньше 1% [24], что не удовлетворяет большое количество пользователей, так как в некоторых высоко масштабируемых системах, например, в центрах обработки данных или поставщиках интернет-услуг количество жестких дисков в одном вычислительном узле может легко достигать тысячи [1]. Отказ каждого из этих жестких дисков может не только увеличить время простоя сервиса, но и привести к потере данных. Задача предсказания отказа жесткого диска является актуальной и для жестких дисков систем хранения данных компании Yadro, которая, согласно данным отчета корпорации IDC (International Data Corporation), является лидером российского рынка внешних систем хранения данных в емкостном выражении с долей 63.7%[1]. На данный момент компания Yadro разрабатывает семейство систем хранения данных Tatlin — Tatlin.Archive, Tatlin.Unified, Tatlin.Object. Так как для систем хранения данных семейства Tatlin прогнозирование отказов жестких дисков не производится, поддержка прогнозирования позволила бы повысить надежность хранения данных в этих системах.

Существующая широко применяемая технология прогнозирования отказа жесткого диска S.M.A.R.T. (*self-monitoring, analysis and reporting technology*) заключается в отслеживании значений набора параметров на предмет превышения соответствующих пороговых значений. Хотя эта технология широко используется, частота предсказания отказов диска с ее помощью является достаточно низкой — ввиду стремления ее создателями снизить количество ложных положительных классификаций, технология S.M.A.R.T. позволяет предсказать от 3% до 10% отказов [15][16][11].

[1] https://st.yadro.com/docs/idc-whitepaper-rus.pdf Дата последнего обращения 27.12.2022

Для повышения частоты предсказания отказа дисков разрабатывались разные подходы, использующие в своей основе данные о поведении диска, собираемые с помощью технологии S.M.A.R.T.. Большинство этих подходов заключается в применении методов машинного обучения к данным, собираемым технологией S.M.A.R.T. [7][15][18][27][28]. Это связано с тем, что задача прогнозирования отказа диска является задачей бинарной классификации, которая в свою очередь является одной из самых распространенных задач, решаемых с помощью алгоритмов машинного обучения.

Ввиду того, что было разработано большое количество подходов для прогнозирования отказов жестких дисков, для поддержки прогнозирования отказов дисков системы хранения данных Talin.Unified целесообразно сначала произвести анализ разработанных подходов. Применение набора подходов для прогнозирования отказа жестких дисков системы хранения данных Tatlin.Unified с последующим сравнением результатов, полученных с помощью этих подходов, позволит выбрать наилучший подход для прогнозирования отказа жестких дисков и интегрировать его, повысив тем самым надежность хранения данных системы.

Постановка задачи

Целью данной работы является поддержать прогнозирование отказов жестких дисков в системе хранения данных Tatlin.Unified.

Для достижения этой цели были поставлены следующие задачи.

- Произвести обзор предметной области;
- Выбрать набор модельных данных и исследовать его, сформулировав ряд закономерностей, характеризующих данные;
- Применить существующие подходы на наборе модельных данных;
- Использовать рассмотренные подходы на реальных данных жестких дисков системы Tatlin.Unified;
- Произвести сравнение примененных к прогнозированию отказов жестких дисков системы Tatlin.Unified подходов;
- Интегрировать наилучший подход к прогнозированию отказов жестких дисков в систему Tatlin.Unified.

1. Существующие подходы к прогнозированию отказа жесткого диска

Существует несколько различных идей, которые могут использоваться для прогнозирования отказа жесткого диска – кластеризация, бинарная классификация, выявление аномалий.

1.1. Прогнозирование отказа жесткого диска с помощью решения задачи бинарной классификации

1.1.1. Использование байесовского классификатора

Применение наивного байесовского классификатора [7] являлось одной из первых попыток применения методов машинного обучения к прогнозированию отказов жестких дисков. В данном исследовании было сделано предположение о том, что отказ жесткого диска невозможно спрогнозировать менее, чем за 48 часов до его отказа. На основе этого предположения класс отказавших жестких дисков был сформирован из записей, относящихся к последним 48 часам наблюдений за параметрами жесткого диска. Применение этого метода дало достаточно низкую точность классификации — доля верных положительных классификаций составила 0.33 при доле 0.001 ложных положительных классификаций. Более поздние исследования применимости этого подхода к прогнозированию отказов жестких дисков дали сходные результаты [8].

Также для прогнозирования отказа жесткого диска использовался древовидный алгоритм Байеса [26], который показал 80% верных положительных классификаций при 3% ложных положительных классификаций. Однако при попытке снизить количество ложных положительных классификаций до 0% частота верных положительных классификаций алгоритма снизилась до 20-30%.

1.1.2. Использование метода опорных векторов

Метод опорных векторов, предложенный в [28], заключается в проецировании исходного набора векторов данных в пространство более высокой размерности и последующем поиске оптимальной разделяющей гиперплоскости [28]. Этот метод также одним из первых стал применяться для прогнозирования отказов дисков [15][16] и позволил классифицировать 50.6% отказов жестких дисков при 0% ложных положительных классификаций. Однако метод требует больших вычислительных ресурсов, и поэтому распространения в предсказывании отказов жестких дисков в режиме реального времени не получил.

1.1.3. Использование скрытой марковской модели

Хотя скрытые марковские модели, предложенные Баумом в 1966 [2], исторически широко применялись для задач распознавания речи, алгоритмы машинного обучения на основе марковских моделей были применены и для прогнозирования отказа жесткого диска [20]. В исследовании [20] авторы рассмотрели последовательности значений параметров жесткого диска, измеренных в последовательные промежутки времени и применили скрытые марковские модели для моделирования этих последовательностей. Подход с использованием скрытой марковской модели позволил достичь 52% положительных классификаций при 0% ложных положительных классификаций, что сравнимо с результатом, полученным с помощью метода опорных векторов.

1.1.4. Метод *K* ближайших соседей

Метод K ближайших соседей, представленный в [5] — это метрический алгоритм классификации, основанный на вычислении оценок сходства между объектами. Для предсказания целевого признака для нового объекта x производятся следующие шаги:

- Вычисляются расстояния от x до всех объектов обучающей выборки;

- Объекты обучающей выборки сортируются по возрастанию расстояний до x;
- Выбираются k объектов с наименьшими расстояниями до x;
- По этим k объектам вычисляется ответ на задачу предсказания.

Данный метод, примененный для решения задачи классификации отказа жесткого диска [19], позволил идентифицировать наибольшее по сравнению с остальным методами количество отказов жестких дисков — 97% при 0.3% ложных положительных классификаций.

1.1.5. Метод случайного леса

Метод случайного леса, заключающийся в использовании набора решающих деревьев для решения задачи классификации, был представлен в [3] и сочетает в себе применение метода бэггинга для построения ансамбля деревьев и метода случайных подпространств.

Примененный для прогнозирования отказов жестких дисков, он показал худшие результаты по сравнению с методом K ближайших соседей, позволив идентифицировать 94.3% отказов дисков при 0.4% ложных положительных классификаций. Тем не менее этот результат является вторым по количеству идентифицированных отказавших дисков среди рассмотренных подходов при одном из самых низких процентов ложных положительных классификаций.

1.2. Прогнозирование отказа жесткого диска с помощью решения задачи выявления аномалий

Хотя с помощью решения соответствующей задачи бинарной классификации для прогнозирования отказа дисков было разработано множество подходов, алгоритмы классификации являются алгоритмами обучения с учителем, то есть, требуют наличие на наборе данных разметки. В случае прогнозирования отказа дисков это означает, что для

всех наблюдений собранных дисковых данных требуется также отметить, было это наблюдение сделано при сломанном диске, или при корректно работающем. В больших корпоративных системах, имеющих множество дисков, определить на лету, сломан тот или иной диск, очень трудно и ресурсоемко, поэтому наборы данных таких систем скорее всего не будут содержать разметку. Отсутствие разметки на реальных данных приводит к необходимости изучения методов машинного обучения, не ориентирующихся на разметку, для прогнозирования дисков. Одним из таких подходов является интерпретация поведения близкого к поломке диска как аномалии, возникшей у корректно работающего диска. В такой интерпретации решение задачи прогнозирования отказа жесткого диска означает решение задачи выявления аномалий среди жестких дисков.

1.2.1. Isolation Forest

Выявление аномалий с помощью алгоритма изолированного леса, представленного в [12], было применено к прогнозированию отказов жестких дисков сравнительно недавно [22]. Алгоритм изолированного леса использует ансамбль деревьев и заключается в рекурсивном разделении набора данных по значению какого-либо случайно выбранного на текущем этапе признака. Так, в результате некоторого числа итераций разделения набора данных, будет образовано дерево, а длина пути от корня до листа, в котором находится тот или иной фрагмент рассматриваемых данных, характеризует степень аномальности этих данных. При случайном рекурсивном разбиении набора данных аномалии как правило имеют более короткие пути от корня дерева к листу, чем остальные элементы (ввиду того, что на определенном шаге алгоритма аномальные наблюдения будут изолированы от остальных по значению какого-либо признака). Алгоритм изолированного леса использует ансамбль деревьев, и в случае, если для ряда деревьев ансамбля какие-то элементы из набора данных имеют маленькое значение длины пути, скорее всего, они будут признаны аномальными. Данный подход показал результаты, значительно превосходящие результаты метода опор-

ных векторов — 84.54% положительных классификаций при 0.0073% ложных положительных классификаций. Также стоит отметить, что хотя доля положительных классификаций у этого метода ниже, чем, например, у алгоритма случайного леса, доля ложных положительных классификаций у этого подхода значительно ниже, чем соответствующая доля ложных положительных классификаций алгоритма случайного леса: 0.0073% против 0.4%.

1.2.2. Использование расстояния Махаланобиса

Подход, использующий расстояние Махаланобиса [18] [30] для прогнозирования отказов жестких дисков, основан на вычислении расстояния Махаланобиса — обобщенном расстоянии, позволяющем измерить сходство между новым наблюдением и набором уже известных наблюдений с помощью рассмотрения корреляций между наблюдениями. Вычисленные значения расстояния Махаланобиса используются четырьмя оценщиками для вычисления значений, сравниваемых с пороговыми. Диск считается отказавшим, если результирующее значение какого-либо из оценщиков превысило пороговое значение.

Данный подход показал результаты, превосходящие метод опорных векторов — 67% положительных классификаций при 0% ложных положительных классификаций. Также подход, использующий расстояние Махаланобиса значительно опередил метод опорных векторов в вычислительной скорости, показав 4.3 минуты против 17983 минут, достигнутых методом опорных векторов. Этот показатель очень важен, так как низкая производительность метода опорных векторов являлась сдерживающим фактором в его использовании для прогнозирования отказов дисков на системах в режиме реального времени. Подход, использующий расстояние Махаланобиса, также позволил предсказать 56% отказов жестких дисков за 20 часов до отказа.

1.2.3. Использование автокодировщиков

Автокодировщики достаточно часто используются для выявления аномалий [14][31][23], и в частности они применяются и для выявления отказов жестких дисков [27]. В исследовании [27] рассмотрено несколько подходов к использованию автокодировщиков для выявления отказов жестких дисков.

Первый из подходов, описанных в [27] использует для классификации наличия отказа диска размер ошибки восстановления — если ошибка восстановления превышает пороговое значение, считается, что произошел отказ жесткого диска. Второй из подходов, описанных в [27] заключается в проверке сходства очередного объекта тренировочной выборки с остальными объектами в пространстве меньшей размерности с использованием расстояния Махаланобиса для измерения сходства. В случае, если расстояние Махаланобиса для очередного объекта превзошло пороговое значение, считается, что произошел отказ жесткого диска. Третий подход, описанный в [27], является комбинацией метрик предыдущих двух подходов. Третий подход показал наибольшую частоту положительных классификаций по сравнению с первыми двумя — 28.09% при 10% ложных положительных классификаций. Результатов применения подходов при 0% ложных положительных классификаций представлено не было.

1.3. Прогнозирование отказа жесткого диска с помощью решения задачи кластеризации

Еще одним потенциально интересным подходом к прогнозированию отказов жестких дисков, не требующим разметки, является кластеризация данных. Кластеризация не применялась к прогнозированию жестких дисков ранее, но высокие результаты кластеризации данных дисков могли бы продемонстрировать наличие определенных моделей поведений дисков в рассматриваемом наборе данных. Низкие результаты кластеризации напротив продемонстрируют невозможность выделить

определенные поведения дисков, и в частности, выделить отказывающие диски.

1.3.1. KMeans

Алгоритм $KMeans$ [13] группирует данные, пытаясь разделить выборки на n групп с одинаковой дисперсией, сводя к минимуму инерцию ($inertia$). Полагая, что $KMeans$ разделяет набор данных из N элементов X на K не пересекающихся кластеров C, каждый из которых описывается средним μ_j элементов кластера, инерцию можно определить следующим образом:

$$\sum_{i=0}^{n} \min \mu_j \in C(\|x_i - \mu_j\|^2) \tag{1}$$

Средние μ_j кластеров также называются центроидами. Недостатком алгоритма $KMeans$ является то, что он требует указание количества кластеров для своей работы, которое влияет на качество полученной кластеризации. Оптимальное количество кластеров может варьироваться в зависимости от набора данных, на котором предполагается проводить кластеризацию.

1.3.2. Bisecting KMeans

Алгоритм $BisectingKMeans$ [4] является итеративным вариантом алгоритма $KMeans$, использующим разделяющую иерархическую кластеризацию. Вместо задания всех центроидов одновременно как это происходит в $KMeans$, центроиды выбираются динамически на основе результатов предыдущей кластеризации: кластер разбивается на два новых кластера до тех пор пока не будет достигнуто целевое количество кластеров.

1.3.3. Mean Shift

Алгоритм $MeanShift$ направлен на обнаружение скоплений в наборе наблюдений равномерной плотности. Это алгоритм, основанный на

итеративной корректировке центроидов таким образом, чтобы они были средним значением точек в заданной области. Затем эти центроиды фильтруются на этапе постобработки, чтобы исключить близкие к дубликатам центроиды и сформировать окончательный набор центроидов.

Положение центроидов корректируется с использованием метода, называемого восхождением на холм, который находит локальные максимумы расчетной плотности вероятности. Пусть на итерации t имеется центроид x, тогда центроид на итерации $t+1$ может быть вычислен следующим образом:

$$x^{t+1} = x^t + m(x^t) \tag{2}$$

где m есть сдвиг среднего значения. Пусть $K(x_i - x)$ — ядерная функция, $N(x)$ — окрестность x, то есть, набор точек, для которых $N(x_i) \neq 0$. Тогда сдвиг среднего значения m задается формулой 3.

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \tag{3}$$

Этот алгоритм не слишком хорошо масштабируется, так как требует поиска множества соседей на каждой итерации в процессе своей работы. В отличие от $KMeans$ и $BisectingKMeans$ не требует задания количества кластеров для проведения кластеризации.

1.3.4. OPTICS

Алгоритм упорядочения точек для обнаружения кластерной структуры (*Ordering points to identify clustering structure, OPTICS*), представленный в [17] кластеризует данные на основе плотности. Он принимает на вход параметр ϵ, а также параметр $MinPts$ — минимальное число точек, необходимое для образования кластера. Точка P называется основной точкой, если в ее окрестности $N_\epsilon(p)$ находятся минимум $MinPts$ точек. Алгоритм состоит из следующей последовательности шагов:

- Выбирается случайная точка P
- Вычисляются все точки, находящиеся в ее $N_\epsilon(p)$ окрестности и кладутся в очередь

- Найденным точкам присваиваются значения достижимости до основной точки и расстояние досягаемости
- В случае, если P — не основная точка, берется следующая точка из очереди и процесс повторяется сначала
- В случае, если P — основная точка, для каждой точки q в $N_\epsilon(p)$ обновляется ее расстояние досягаемости до основной точки. Все непосещенные ранее точки q складываются в очередь для дальнейшей обработки.
- В результате посещения всех точек алгоритмом получается набор расстояний досягаемости, на основании которых данные кластеризуются.

Данный алгоритм очень хорошо масштабируется и не требует указания количества кластеров для своей работы, поэтому может представить хорошие результаты при работе с таким большим набором данных, как множество параметров различных дисков.

1.3.5. Spectral Clustering

Алгоритм спектральной кластеризации, изложенный в [32], заключается в применении стандартных алгоритмов кластеризации (например, $KMeans$) на данных сниженной размерности. Алгоритм получает на вход некоторую матрицу сходства S размерности $\mathbb{R}^{n*n}$ и целевое количество кластеров k и состоит из следующей последовательности шагов:

- Построить на основе полученной матрицы сходства граф сходства
- Вычислить для построенного графа дискретный оператор Лапласа L
- Вычислить k собственных векторов $u_1, \dots u_k$ из $Lu = \lambda Du$
- Составить матрицу $U \in \mathbb{R}^{n*k}$, где столбцами являются векторы $u_1, \dots u_k$

- Считая $y_i \in \mathbb{R}^k$ для $i = 1 \dots n$ строками матрицы U, провести кластеризацию для $y_{i_{i=1\dots n}} \in \mathbb{R}^k$ с помощью алгоритма кластеризации $(KMeans)$ на кластера $C_1, \dots C_k$

- Выдать в качестве результата набор кластеров $A_1, \dots A_k$, где $A_i = j | y_j \in C_k$

Алгоритм спектральной кластеризации требует указания количества кластеров, что усложняет его использование, так как оптимальное количество кластеров в кластеризации сильно зависит от набора данных.

2. Применение подходов на наборе модельных данных

2.1. Выбор набора модельных данных

В большинстве исследований, приведенных в главе 1, применимость разработанных подходов к прогнозированию отказов жестких дисков оценивалась на каком-либо из следующих наборов данных [21]:

- *University of California dataset*[2]
- *Baidu dataset*
- *Quantum Corporation dataset*
- *Backblaze dataset*[3]

Набор данных, предоставленный университетом Калифорнии использовался, например, в работах [15][16]. Недостатком этого набора данных является то, что его признаки не являются непосредственно S.M.A.R.T. параметрами, что снижает применимость выводов, сделанных на основе этого набора данных, к реальным данным, составленным из значений S.M.A.R.T.. Кроме того, набор данных университета Калифорнии был собран достаточно давно и поэтому гипотезы, сделанные на основе этих данных, могут быть не применимы к данным, собранным в настоящее время.

Набор данных корпорации *Quantum* использовался, например, в работе [7]. Так как этот набор данных имеет 11 признаков, соответствующих S.M.A.R.T. параметрам и был собран достаточно давно, гипотезы, опробованные на этом наборе данных, могут быть не применимы к реальным данным жестких дисков, имеющих более 200 S.M.A.R.T. параметров.

[2] https://web.archive.org/web/20100611213812/http://cmrr.ucsd.edu/ Дата последнего обращения 27.12.2022

[3] https://www.backblaze.com/b2/hard-drive-test-data.html Дата последнего обращения 27.12.2022

Из двух аналогичных наборов данных, предоставленных центрами обработки данных *Baidu* и *Backblaze*, был выбран предобработанный набор данных[4] на основе данных центра *Backblaze*. Этот набор данных содержит данные одной модели жесткого диска Seagate ST4000DM000 и содержит данные о 120 днях работы до отказа для отказавших дисков и 120 произвольных днях работы для дисков, работающих корректно.

Этот набор данных содержал дублирующие друг друга признаки, а также признаки, имеющие константные значения на всем наборе данных, что позволило в результате предобработки сократить количество признаков с 90 до 17.

2.2. Применение алгоритмов бинарной классификации

2.2.1. Гипотеза о важности S.M.A.R.T. параметров при классификации отказа диска

Следующие S.M.A.R.T. параметры считаются критически важными для прогнозирования отказа дисков [5]

- Количество перераспределенных секторов (*Reallocated sectors count*)
- Сквозная ошибка (*End-to-end error*)
- Сообщения о неисправимых ошибках (*Reported Uncorrectable Errors*)
- Текущее количество ожидающих секторов (*Current Pending Sector Count*)
- Количество неисправимых секторов (*Uncorrectable Sector Count*)

Исходя из того, что эти параметры считаются критическими, использование только их для обучения может улучшить характеристики

[4] https://www.kaggle.com/datasets/awant08/hard-drive-failure-prediction-st4000dm000 Дата последнего обращения 27.12.2022

[5] https://en.wikipedia.org/wiki/Self-Monitoring,_Analysis_and_Reporting_Technology Дата последнего обращения 27.12.2022

модели. Справедливость данного предположения проверялась с помощью нескольких методов: метода опорных векторов, метода K ближайших соседей, метода случайного леса. Результаты работы методов измерялись с помощью метрик FAR — частоты ложных положительных классификаций и FDR (*Failure Detection Rate*) — отношения выявленных отказавших дисков к общему числу отказавших дисков.

2.2.2. Применение метода опорных векторов

Так как рассмотренный в разделе 1.1.2 метод опорных векторов одним из первых применялся к прогнозированию отказа жестких дисков, было принято решение оценить применимость этого метода к прогнозированию отказов жестких дисков модели Seagate ST4000DM000. Данный метод подтвердил свою низкую производительность в прогнозировании отказа жестких дисков и на 17 выделенных признаках за время, значительно превосходящее время обучения других моделей, не удалось обучить модель с использованием метода опорных векторов.

В результате применения данного метода к данным модели Seagate ST4000DM000 на 5 признаках, соответствующих критическим параметрам S.M.A.R.T., были достигнуты следующие метрики:

- $FDR = 1 - \dfrac{2812}{(2812 + 395)} = 0.123$
- $FAR = \dfrac{67}{95020} = 0.0007$

2.2.3. Применение метода K ближайших соседей

Метод K ближайших соседей был выбран ввиду того, что показал наибольшую точность 0.974% при 0.003% ложных положительных классификаций [21][19], а также не требовал предварительной подготовки для его использования, облегчая процесс проверки гипотезы.

В результате применения данного метода к данным модели Seagate ST4000DM000на 17 выделенных признаках были достигнуты следующие метрики:

- $FDR = 1 - \frac{740}{(740 + 2467)} = 0.769$
- $FAR = \frac{411}{95020} = 0.004$

В результате применения данного метода к данным модели Seagate ST4000DM000 на 5 признаках, соответствующих критическим S.M.A.R.T., были достигнуты следующие метрики:

- $FDR = 1 - \frac{2817}{(2817 + 390)} = 0.122$
- $FAR = \frac{64}{95020} = 0.0006$

2.2.4. Применение метода случайного леса

Метод случайного леса был выбран потому, что показал один из лучших результатов [21][19], классифицировав 0.943% отказов жестких дисков при 0.004 ложных положительных классификаций,

В результате применения данного метода к данным модели Seagate ST4000DM000на 17 выделенных признаках были достигнуты следующие метрики:

- $FDR = 1 - \frac{586}{(586 + 2621)} = 1 - 0.183 = 0.817$
- $FAR = \frac{212}{95020} = 0.002$

В результате применения данного метода к данным модели Seagate ST4000DM000 на 5 признаках, соответствующих критическим S.M.A.R.T., были достигнуты следующие метрики:

- $FDR = 1 - \frac{2808}{(2808 + 399)} = 0.124$
- $FAR = \frac{67}{95020} = 0.0007$

Использование при обучении только признаков, соответствующих S.M.A.R.T. параметрам, считающимся критическими, значительно увеличило количество ложных отрицательных классификаций, таким образом ухудшив характеристики модели. Таким образом, предположение о том, что использование только признаков, соответствующих критическим S.M.A.R.T. параметрам, улучшит показатели моделей, было опровергнуто, и в дальнейшем при обучении предполагается использовать остальные S.M.A.R.T. параметры.

2.3. Применение алгоритмов выявления аномалий

Для оценки работы алгоритмов выявления аномалий используются те же метрики, что и для оценки работы алгоритмов бинарной классификации, а именно *FAR* — частота ложных положительных классификаций и *FDR* (*Failure Detection Rate*) — отношение выявленных отказавших дисков к общему числу отказавших дисков.

2.3.1. Isolation Forest

Алгоритм изолированного леса применялся в соответствии с проведенным исследованием о прогнозировании отказа дисков [22]. Было создано две модели, использующие алгоритм изолированного леса со значениями параметра *contamination* (равному доле наблюдений в наборе данных, которые следует идентифицировать как аномальные) равными 0.01 и 0.0002. В результате применения алгоритма изолированного леса со значением параметра $contamination = 0.01$ к данным были достигнуты следующие значения метрик:

- $FAR = \dfrac{886}{121910 + 886} = 0.0072$
- $FDR = 1 - \dfrac{3654}{4162 + 2} = 0.1225$

В результате применения алгоритма изолированного леса с $contamination = 0.0002$ к данным были достигнуты следующие значения метрик:

- $FDR = 1 - \frac{4162}{4162 + 2} = 0.0004$
- $FAR = \frac{35}{121910 + 35} = 0.0002$

Таким образом, применение изолированного леса со значением параметра $contamination = 0.01$ позволило достичь более низкой частоты ложных классификаций (FAR) при более высокой частоте детекции отказавших дисков (FDR) по сравнению по сравнению с результатами применения метода случайного леса, изложенными в 2. Дальнейшее повышение параметра *contamination* позволит достичь большего значения частоты классификации отказавших дисков.

2.3.2. Расстояние Махаланобиса

Алгоритм выявления аномалий с помощью расстояния Махаланобиса был реализован в соответствии с исследованием [18] [30]. В результате выявления аномалий с помощью расстояния Махаланобиса были достигнуты следующие метрики:

- $FDR = 1 - \frac{3970}{3970 + 194} = 0.0466$
- $FAR = \frac{712}{122084 + 712} = 0.0058$

Так как алгоритм использует 4 оценщика, зависящих от заданных параметров, в процессе своей работы, результаты работы алгоритма зависят как от набора данных, так и от того, насколько оптимально подобраны заданные параметры. Потенциально текущие результаты могут быть улучшены, так как подбор оптимальных параметров оценщиков не проводился.

2.4. Применение алгоритмов кластеризации

2.4.1. Применение метода главных компонент для сокращения размерности набора данных

Работа алгоритмов кластеризации требует много ресурсов, поэтому применение их напрямую к набору данных жестких дисков, содержа-

щему множество признаков, может занять очень большое количество времени. Так как рассматриваемый набор данных жестких дисков даже после очистки содержит большое количество признаков, для применения алгоритмов кластеризации было полезно сократить размерность данных. Для этого использовался метод главных компонент [25], являющийся одним из наиболее распространенных методов для снижения размерности данных. Результаты применения метода главных компонент, представленные в таблице 2.4.1 показали, что снижение размерности данных до 4 позволит сохранить 80% информации о наборе данных дисков, снижение размерности данных до 7 позволит сохранить 95% информации соответственно.

Размерность данных	Количество информации
4	80%
7	95%

К набору данных сниженной размерности было применено несколько алгоритмов кластеризации, описанных в 1. Для сравнения алгоритмов кластеризации используются метрики, отличные от метрик, использующихся для оценки решений задачи классификации. Так как набор данных обладает разметкой, для повышения точности измерения работы различных алгоритмов кластеризации использовались метрики, учитывающие наличие разметки: скорректированный индекс Рэнда (ARI) [10], скорректированная взаимная информация (AMI) [29], индекс Фаулкса-Мэллоуза (FMI) [6].

Определение 2.4.1 *Скорректированный индекс Рэнда (ARI) — метрика, отражающая меру подобия между двумя кластерами. Пусть C — кластеризация, заданная разметкой, K — кластеризация, полученная с помощью алгоритма, a — количество пар элементов, которые попали в один кластер кластеризации C и в один кластер кластеризации K, b — количество пар элементов, которые попали в разные кластера кластеризации C и в разные кластера кластеризации K . Тогда скорректированный индекс Рэнда может быть выражен следу-*

ющей формулой:

$$ARI = \frac{RI - \mathbb{E}[RI]}{max(RI) - \mathbb{E}[RI]} \tag{4}$$

RI есть нескорректированный индекс Рэнда, который выражается следующим образом:

$$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{5}$$

где $C_2^{n_{samples}}$ — число всевозможных пар в рассматриваемом наборе данных.

Определение 2.4.2 *Пусть U, V есть две разметки на N объектах. Их энтропия определяется следующим образом:*

$$H(U) = -\sum_{i=1}^{|U|} P(i) * \log P(i) \tag{6}$$

$$H(V) = -\sum_{j=1}^{|V|} P'(j) * \log P'(j) \tag{7}$$

Их взаимная информация может быть выражена следующей формулой:

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|U|} \frac{U_i \cap V_j}{N} \log(\frac{N|U_I \cap U_j|}{|U_i||V_j|}) \tag{8}$$

и математическое ожидание взаимной информации может быть выражено следующим образом:

$$\begin{gathered}\mathbb{E}[MI] = \sum_{i=1}^{|U|}\sum_{j=1}^{|U|}\sum_{n_{ij}=(a_i+b_j-N)^+}^{min(a_i,b_j)} \frac{n_{ij}}{N}\log(\frac{Nn_{ij}}{a_i b_j}) * \\ * \frac{a_i! b_j! (N-a_i)!(N-b_j)!}{N! n_{ij}! (a_i - n_{ij})!(b_j - n_{ij})!(N - a_i - b_j + n_{ij})!}\end{gathered} \tag{9}$$

где $a_i = |U_i|$ и $b_j = |V_j|$ (число элементов в U_i и в V_j соответственно). Тогда скорректированная взаимная информация может быть выражена следующим образом:

$$AMI = \frac{MI - \mathbb{E}[MI]}{mean(H(U), H(V) - \mathbb{E}[MI]} \tag{10}$$

Определение 2.4.3 *Индекс Фаулкса-Мэллоуза определяется как среднее геометрическое точности и полноты и определяется следующим образом:*

$$FMI = \frac{TP}{sqrt((TP + FP)(TP + FN))} \tag{11}$$

где TP (True Positive) — число пар точек, которые принадлежат к одним и тем же кластерам для обеих кластеризаций, FP (False Positive) — число пар точек, которые принадлежат к одним и тем же кластерам для кластеризации, задаваемой разметкой, и принадлежат к разным кластерам для предсказанной кластеризации, FN (False Negative) — число пар точек, которые принадлежат к разным кластерам для кластеризации, задаваемой разметки и к одному кластеру в случае предсказанной кластеризации.

2.4.2. Сравнительный анализ алгоритмов кластеризации

Результаты применения алгоритмов кластеризации *KMeans*, *Bisecting KMeans*, *Spectral Clustering*, *Mean Shift*, *OPTICS* на наборе данных сниженных размерностей 4 и 7 представлены в таблице 2.4.2.

Алгоритм	Размерность данных	Время обучения, сек	ARI	AMI	FMI
KMeans	4	1.71194	0.00071	0.00233	0.68589
KMeans	7	1.80218	0.00093	0.00240	0.68586
OPTICS	4	3170.495	0.00034	0.01124	0.31021
OPTICS	7	4749.34209	-0.00044	0.00794	0.30463
MeanShift	4	12067.55847	0.0	0.0	0.96972
MeanShift	7	21211.85286	0.00328	0.00383	0.96162
SpectralClustering	4	32619.47111	0.00493	0.00266	0.00534
SpectralClustering	7	25367.72637	0.00493	0.00053	0.93507

Для измерения качества кластеризации использованы приведенные в 2.4.1 2.4.2 2.4.3 метрики ARI, AMI, FMI. В результате анализа таблицы можно сделать вывод, что проведение кластеризации для рассмат-

риваемого набора данных показало плохие результаты для всех алгоритмов — так, например, максимальное значение ARI не превышает 0.005, что свидетельствует о том, что метки кластеров расставлены на данных почти случайным образом. Это демонстрирует то, что из рассмотренных данных жестких дисков не удалось выделить некоторые общие модели поведения жестких дисков.

2.5. Анализ примененных подходов

В исследовании были рассмотрены три подхода к прогнозированию отказа жестких дисков: прогнозирование отказа жесткого диска с помощью решения задачи бинарной классификации, с помощью решения задачи выявления аномалий, прогнозирование с помощью кластеризации. Недостатком примененных подходов, основывающихся на решении задачи бинарной классификации, является необходимость разметки набора данных, которую трудно получить для реальных данных жестких дисков, собранных с корпоративных систем. Подходы к прогнозированию, основывающиеся на выявлении аномалий, не требуют разметки. В исследовании они показали более низкие доли детекции отказавших дисков по сравнению с алгоритмами бинарной классификации. Доли ложных классификаций у этих алгоритмов также значительно ниже алгоритмов бинарной классификации, поэтому эти алгоритмы потенциально позволят классифицировать большее количество отказавших дисков при соответствующем повышении частот ложных классификаций. Также для прогнозирования отказа жестких дисков были применены различные алгоритмы кластеризации. Применение алгоритмов кластеризации показало низкие результаты, что может говорить либо о недостаточности данных, либо о том, что для рассматриваемого набора данных нельзя выделить несколько явных моделей поведения, на которые в дальнейшем можно было бы кластеризовать диски. Таким образом, для рассматриваемого набора данных с точки зрения применимости на реальных системах наиболее перспективным оказался подход прогнозирования отказа дисков с помощью выявления аномалий.

Заключение

В ходе учебной практики были достигнуты следующие задачи:

- Проведен обзор предметной области и рассмотрены алгоритмы, реализующие такие подходы машинного обучения для прогнозирования отказа жесткого диска как:
  - Прогнозирование отказа жесткого диска с помощью решения задачи бинарной классификации;
  - Прогнозирование отказа жесткого диска с помощью решения задачи выявления аномалий;
  - Прогнозирование отказа жесткого диска с помощью кластеризации.
- Для анализа модельного набора данных применены следующие методы, рассмотренные в обзорной части работы:
  - Метод опорных векторов, случайного леса и K ближайших соседей в рамках решения задачи бинарной классификации;
  - Метод изолированного леса, выявления аномалий с использованием расстояния Махаланобиса в рамках решения задачи выявления аномалий;
  - Методы *KMeans*, *Bisecting KMeans*, *Spectral Clustering*, *Mean Shift*, *OPTICS* в рамках решения задачи кластеризации.
- Результаты применения методов проанализированы, сделан вывод о том, что наиболее перспективными для прогнозирования дисков на данный момент являются методы выявления аномалий.

В рамках продолжения исследований планируется выполнить следующие задачи:

- Оценить применимость еще не рассмотренных подходов на модельных данных — иерархических алгоритмов кластеризации, а также новых алгоритмов выявлений аномалий;

- Использовать рассмотренные подходы на реальных данных жестких дисков системы Tatlin.Unified.
- Произвести сравнение примененных к прогнозированию отказов жестких дисков системы Tatlin.Unified подходов.

Список литературы

[1] Jiang Weihang, Hu Chongfeng, Zhou Yuanyuan, and Kanevsky Arkady. Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics // ACM Trans. Storage. — 2008. — nov. — Vol. 4, no. 3. — Access mode: https://doi.org/10.1145/1416944.1416946.

[2] Baum Leonard E. and Petrie Ted. Statistical Inference for Probabilistic Functions of Finite State Markov Chains // Annals of Mathematical Statistics. — 1966. — Vol. 37. — P. 1554–1563.

[3] Breiman L. Random Forests // Machine Learning. — 2001. — 10. — Vol. 45. — P. 5–32.

[4] Di Jian and Gou Xinyu. Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization // J. Comput. — 2018. — Vol. 13. — P. 588–595.

[5] Fix Evelyn and Hodges Joseph L. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties // International Statistical Review. — 1989. — Vol. 57. — P. 238.

[6] Fowlkes E. B. and Mallows C. L. A Method for Comparing Two Hierarchical Clusterings // Journal of the American Statistical Association. — 1983. — Vol. 78, no. 383. — P. 553–569. — https://www.tandfonline.com/doi/pdf/10.1080/01621459.1983.10478008.

[7] Hamerly Greg and Elkan Charles. Bayesian Approaches to Failure Prediction for Disk Drives // Proceedings of the Eighteenth International Conference on Machine Learning. — San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. — 2001. — ICML '01. — P. 202–209.

[8] Tomer Vikas, Sharma Vedna, Gupta Sonali, and Singh Devesh. Hard disk drive failure prediction using SMART attribute // Materials Today: Proceedings. — 2021. — 04. — Vol. 46.

[9] Hilbert Martin and López Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information // Science (New York, N.Y.). — 2011. — 02. — Vol. 332. — P. 60–5.

[10] Hubert Lawrence and Arabie Phipps. Comparing partitions // Journal of Classification. — 1985. — Dec. — Vol. 2, no. 1. — P. 193–218. — Access mode: https://doi.org/10.1007/BF01908075.

[11] Hughes Gordon, Murray Joseph, Kreutz-Delgado Ken, and Elkan Charles. Improved disk-drive failure warnings // Reliability, IEEE Transactions on. — 2002. — 10. — Vol. 51. — P. 350 – 357.

[12] Liu Fei Tony, Ting Kai Ming, and Zhou Zhi-Hua. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. — 2008. — P. 413–422.

[13] Lloyd Stuart P. Least squares quantization in PCM // IEEE Trans. Inf. Theory. — 1982. — Vol. 28. — P. 129–136.

[14] Morales-Forero A. and Bassetto S. Case Study: A Semi-Supervised Methodology for Anomaly Detection and Diagnosis // 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). — 2019. — P. 1031–1037.

[15] Murray Joseph, Hughes Gordon, and Kreutz-Delgado Ken. Hard drive failure prediction using non-parametric statistical methods. — 2003. — 01.

[16] Murray Joseph F., Hughes Gordon F., and Kreutz-Delgado Kenneth. Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application // J. Mach. Learn. Res. — 2005. — dec. — Vol. 6. — P. 783–816.

[17] Ankerst Mihael, Breunig Markus M., Kriegel Hans-Peter, and Sander Jörg. OPTICS: Ordering Points to Identify the Clustering Structure // Proceedings of the 1999 ACM SIGMOD International

Conference on Management of Data. — New York, NY, USA : Association for Computing Machinery. — 1999. — SIGMOD '99. — P. 49–60. — Access mode: https://doi.org/10.1145/304182.304187.

[18] Wang Yu, Miao Qiang, Ma Eden W. M., Tsui Kwok-Leung, and Pecht Michael G. Online Anomaly Detection for Hard Disk Drives Based on Mahalanobis Distance // IEEE Transactions on Reliability. — 2013. — Vol. 62, no. 1. — P. 136–145.

[19] Pitakrat Teerat, van Hoorn André, and Grunske Lars. A comparison of machine learning algorithms for proactive hard disk drive failure detection. — 2013. — 06. — P. 1–10.

[20] Zhao Ying, Liu Xiang, Gan Siqing, and Zheng Weimin. Predicting disk failures with HMM- and HSMM-based approaches. — 2010. — 07. — Vol. 6171. — P. 390–404.

[21] Garcia Marco, Ivanov Vladimir, Kozar Anastasia, Litvinov Stanislav, Reznik Alexey, Romanov Vitaly, and Succi Giancarlo. Review of techniques for predicting hard drive failure with SMART attributes // International Journal of Machine Intelligence and Sensory Signal Processing. — 2018. — 01. — Vol. 2. — P. 151.

[22] Rombach Philipp and Keuper Janis. SmartPred: Unsupervised Hard Disk Failure Detection // High Performance Computing: ISC High Performance 2020 International Workshops, Frankfurt, Germany, June 21–25, 2020, Revised Selected Papers. — Berlin, Heidelberg : Springer-Verlag. — 2020. — P. 235–246. — Access mode: https://doi.org/10.1007/978-3-030-59851-8_15.

[23] Sakurada Mayu and Yairi Takehisa. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction // Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. — New York, NY, USA : Association for Computing Machinery. — 2014. — MLSDA'14. — P. 4–11. — Access mode: https://doi.org/10.1145/2689746.2689747.

[24] Schroeder Bianca and Gibson Garth A. Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? // Proceedings of the 5th USENIX Conference on File and Storage Technologies. — USA : USENIX Association. — 2007. — FAST '07. — P. 1–es.

[25] Shlens Jonathon. A Tutorial on Principal Component Analysis. — 2014. — 1404.1100.

[26] Tan Yongmin and Gu Xiaohui. On Predictability of System Anomalies in Real World // 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. — 2010. — P. 133–140.

[27] Pereira Francisco Lucas F., Castro Chaves Iago, Gomes João Paulo P., and Machado Javam C. Using Autoencoders for Anomaly Detection in Hard Disk Drives // 2020 International Joint Conference on Neural Networks (IJCNN). — 2020. — P. 1–7.

[28] Vapnik Vladimir. The Nature of Statistical Learning Theory. — 2000. — 01. — P. 69–91. — ISBN: 978-1-4419-3160-3.

[29] Vinh Nguyen Xuan, Epps Julien, and Bailey James. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? // Proceedings of the 26th Annual International Conference on Machine Learning. — New York, NY, USA : Association for Computing Machinery. — 2009. — ICML '09. — P. 1073–1080. — Access mode: https://doi.org/10.1145/1553374.1553511.

[30] Wang Yu, Miao Qiang, and Pecht Michael. Health monitoring of hard disk drive based on Mahalanobis distance // 2011 Prognostics and System Health Managment Confernece. — 2011. — P. 1–8.

[31] Zhou Chong and Paffenroth Randy C. Anomaly Detection with Robust Deep Autoencoders // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA : Association for Computing Machinery. — 2017. —

KDD '17. — P. 665–674. — Access mode: https://doi.org/10.1145/3097983.3098052.

[32] von Luxburg Ulrike. A Tutorial on Spectral Clustering. — 2007. — 0711.0189.