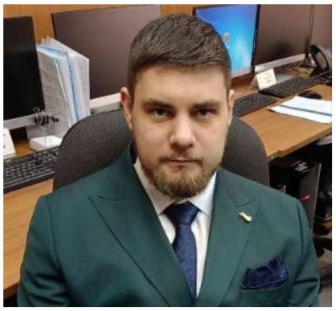
УДК 658.8:004.9

ФОРМИРОВАНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ ПРЕДИКТИВНОГО АНАЛИЗА СОСТОЯНИЯ ДИСКОВЫХ ПОДСИСТЕМ НА ОСНОВЕ АТРИБУТОВ S.M.A.R.T.



М.С. Борисенков
Аспирант ТУ им. А.А. Леонова (филиал) МИИГАиК
borisenkov.matvey@mail.ru

С.Н. Шульженко
Профессор ТУ им. А.А. Леонова (филиал) МИИГАиК, доктор технических наук
shulzh79@mail.ru

М.С. Борисенков

Окончил Технологический университет имени А.А. Леонова.. Область научных интересов связана с ракетокосмической областью и разработкой информационно-прогностических систем.

С.Н. Шульженко

Окончил Тульский государственный университет. Область научных интересов связана с разработкой и совершенствованием интеллектуальных информационных систем и их применением в производственных задачах

Аннотация. В статье предложена методика оптимизации признакового пространства телеметрии серверного оборудования на основе корреляционного анализа и аппарата информационной энтропии.

Ключевые слова: прогнозное обслуживание, телеметрия серверного оборудования, препроцессинг данных, коэффициент корреляции Спирмена, прирост информации, энтропия Шеннона, снижение размерности признаков.

Введение и постановка задачи. Экспоненциальное наращивание объемов обрабатываемой информации в сфере высокопроизводительных вычислений и облачных технологий выдвигает на первый план проблему бесперебойности работы ИТ-комплексов. Существующие реактивные подходы к обслуживанию, когда ремонт инициируется лишь после фактического отказа компонента, демонстрируют свою неэффективность в масштабах современных дата-центров. Подобная практика неизбежно влечет за собой рост эксплуатационных расходов и регулярные сбои в выполнении условий соглашений об уровне сервиса (SLA) [3]. В этом контексте переход к предиктивной модели обслуживания (Predictive Maintenance), нацеленной на прогнозирование состояния серверных платформ и систем хранения, приобретает характер первостепенной научно-прикладной проблемы.

Хотя современные аппаратные платформы предоставляют развитый инструментарий для сбора телеметрии (SMART, IPMI, SNMP), непосредственное применение всего объема регистрируемых параметров при создании предиктивных моделей наталкивается на фундаментальные трудности. Проблема заключается в чрезмерной размерности и зашумленности исходного массива данных. Пространство признаков оказывается избыточным, что не просто усложняет вычислительные процедуры обучения нейросетей, но и провоцирует их переобучение. В итоге это негативно сказывается на достоверности долгосрочных прогнозов.

Настоящее исследование направлено на создание и теоретическое обоснование методики, позволяющей сформировать оптимальный вектор информативных признаков. В основе подхода лежит математический анализ параметров, характеризующих деградационные процессы. Объектом изучения служат массивы телеметрических временных рядов, которые регистрируются в ходе эксплуатации серверных узлов и дисковых подсистем. Реализация данной цели обеспечит существенное снижение размерности исходных данных, подаваемых на вход интеллектуальных прогнозных моделей. При этом ключевым условием остается сохранение физической интерпретируемости процессов износа оборудования.

Математический аппарат анализа и фильтрации признаков. Переход от теоретической формулировки задачи к её практическому воплощению в виде интеллектуальной модели диктует необходимость строгого обоснования отбора входных параметров. В рамках данного исследования эта процедура реализуется через последовательное применение аппарата математической статистики и положений теории информации. Такой подход обеспечивает формирование компактного, но при этом предельно насыщенного информацией признакового пространства. Процесс обработки данных начинается с этапа предварительной фильтрации и нормализации, направленного на исключение параметров, не обладающих предсказательной способностью. Первоначально из рассмотрения изымаются квазипостоянные признаки, дисперсия которых 2 стремится к нулю, так как отсутствие вариативности в данных делает невозможным фиксацию динамики деградации. Для преодоления проблемы разнородности шкал телеметрических метрик (температурных показателей, счетчиков ошибок и коэффициентов загрузки) применяется процедура Z-масштабирования:

$$z = \frac{x-\mu}{\sigma},$$

где – математическое ожидание, а – среднеквадратичное отклонение признака кореляции [1]. Такое преобразование приводит все анализируемые величины к единому масштабу, обеспечивая корректную работу последующих алгоритмов.

Дальнейшее уточнение признакового пространства осуществляется путем анализа статистических взаимосвязей для выявления избыточных параметров. С этой целью рассчитывается коэффициент ранговой корреляции Спирмена который, в отличие от линейных методов, позволяет эффективно идентифицировать монотонные нелинейные зависимости, характерные для процессов постепенного износа аппаратных компонентов:

$$\rho = 1 - \frac{6\sum_1^2 * d}{n(n^2-1)}$$

При обнаружении высокой степени корреляции (ρ> 0,85) между парой признаков один из них подлежит исключению. Это позволяет устранить явление мультиколлинеарности, которое негативно влияет на стабильность прогнозных моделей. Завершающим этапом математического обоснования является ранжирование оставшихся параметров по их информационной значимости. В основу данного подхода положен расчет прироста информации (Information Gain), базирующийся на определении энтропии Шеннона HS, отражающей меру неопределенности состояния системы (нормальное функционирование или предаварийный режим) [2]:

$$H(S) = -p_i$$

Информационная ценность каждого конкретного признака A оценивается через снижение энтропии системы после включения данного признака в анализ:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_V|}{|S|} * H(S_V)$$

Использование иерархической методологии открывает возможность для строгой математической верификации результирующего вектора признаков. Этот верифицированный вектор, в свою очередь, формирует фундаментальную базу для последующей экспериментальной проверки.

Экспериментальное исследование и анализ результатов. Предложенный математический аппарат прошел экспериментальную проверку на обезличенных телеметрических данных, собранных за год непрерывной работы серверного кластера и систем хранения. Практическая часть исследования подтвердила теоретические выкладки и дала возможность измерить результативность созданной методики формирования признакового пространства. В качестве исходного набора данных был сформирован массив, включающий 42 первичных телеметрических параметра. На этапе предварительной обработки, следуя описанному ранее математическому алгоритму, была проведена фильтрация признаков с низкой дисперсией (VarX<0.01), что позволило исключить статические идентификаторы и параметры конфигурации, не изменяющиеся во времени и не несущие прогностической ценности. Последующее применение Z-масштабирования обеспечило нормализацию оставшихся 18 метрик, устранив влияние различий в единицах измерения и подготовив данные для корреляционного анализа. Центральным элементом анализа статистических взаимосвязей выступило построение матрицы ранговой корреляции Спирмена. Выбор данного метода позволил выявить устойчивые зависимости, которые в условиях зашумленности лог-файлов серверного оборудования часто принимают нелинейный характер. На представленной ниже тепловой карте (рис. 1) визуализированы коэффициенты между ключевыми метриками.

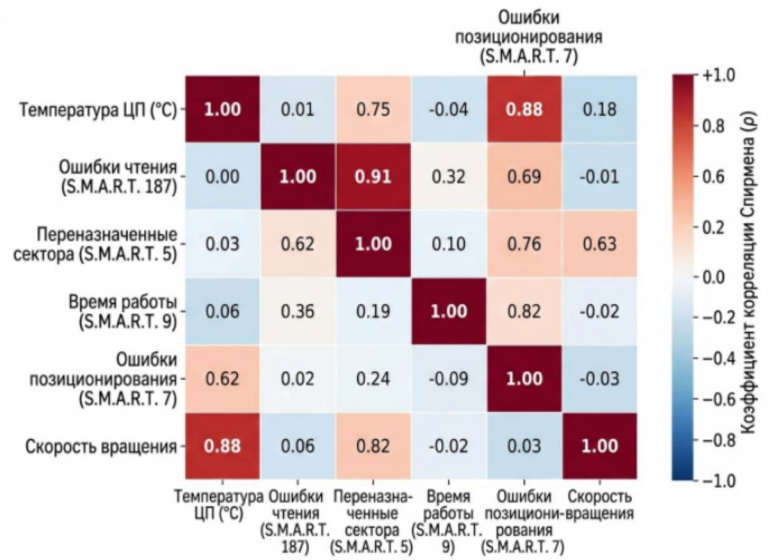


Рисунок 1. Тепловая карта ранговой корреляции Спирмена для параметров телеметрии серверного оборудования

Выявление кластеров с коэффициентом $\rho>0,85$ (например, между показателями «Reallocated Sectors Count» и «Current Pending Sector Count») послужило математическим основанием для исключения одного из дублирующих признаков, что критически важно для предотвращения переобучения будущей интеллектуальной модели.

Для последующего упорядочивания признаков был задействован метод, основанный на концепции информационной энтропии. Оценивая величину прироста информации (Information Gain), удалось выявить те параметры, что в максимальной степени уменьшают энтропийную неопределенность в оценке текущего состояния технической системы. В таблице 1 представлены результаты вычислений, где значения IG ранжированы по убыванию, что позволяет выделить ядро наиболее информативных переменных.

Таблица 1. Математические показатели значимости признаков в прогнозировании отказов

Признак (Параметр телеметрии)	**Коэффициент вариации**	**Значение Information Gain (IG)**	**Принятое решение**
Reallocated Sector Count	0.84	0.428	Включить в модель
Average CPU Core Temp	0.32	0.315	Включить в модель
Reported Uncorrectable Errors	0.91	0.284	Включить в модель
Read Error Rate (Raw)	0.76	0.192	Включить в модель
Power-On Hours	0.05	0.156	Исключить (низкая динамика)
Spin-Up Time	0.12	0.082	Исключить (низкий IG)

Физическая обоснованность математических выводов проверялась через сопоставление временных рядов (риунок 2). Графики, иллюстрирующие динамику ключевых параметров за 72 часа до критического отказа, отчетливо выявляют лавинообразную деградацию системы. Это проявляется, например, в синхронном нарастании температуры центрального процессора и частоты ошибок при чтении данных – явление, которое находит свое математическое выражение в скачкообразном изменении градиента функции состояния.

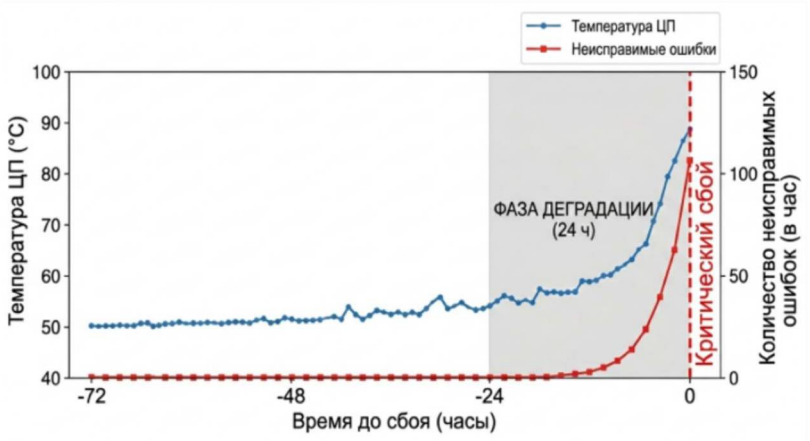


Рисунок 2. Временная корреляция ключевых параметров деградации (температуры ЦП и неисправимых ошибок)

Итогом экспериментального исследования стало формирование оптимизированного вектора из 7 информативных признаков. Математически доказано, что сокращение размерности признакового пространства на 83% не привело к существенной потере информационной ценности (суммарный IG отобранных признаков покрывает более 90% энтропии целевой переменной). Таким образом, сформированная база данных является репрезентативной и достаточной для реализации следующего этапа работы – построения интеллектуальной модели прогнозирования на основе глубокого обучения.

Заключение. В данной статье представлена и экспериментально проверена методика построения оптимального признакового пространства для систем прогнозирования технического состояния серверов и систем хранения данных. В отличие от традиционных эмпирических методов отбора параметров, предложенный алгоритм использует формальный математический аппарат, основанный на принципах теории информации и корреляционного анализа.

Основные результаты исследования:

1. Предлагаемый двухэтапный механизм фильтрации реализует последовательную обработку данных: на первом этапе применяется Z-масштабирование, затем устраняется мультиколлинеарность признаков с использованием коэффициента Спирмена, а завершающим шагом становится их ранжирование по метрике прироста информации (Information Gain).

2. Экспериментальные результаты демонстрируют существенную информативность динамических параметров, таких как температурные режимы и отдельные S.M.A.R.T.-счетчики. Их совокупный вклад в общую энтропию модели позволил сократить размерность входного вектора на 83%, не нарушая репрезентативности исходной выборки.

3. Для разработки прогнозных моделей, основанных на глубоком обучении, была сформирована соответствующая информационная база. Математически верифицированный набор, включающий семь ключевых признаков, позволяет существенно сократить вычислительные затраты на этапе обучения нейронных сетей, сохраняя при этом высокую точность прогнозирования критических отказов.

Предложенная методология предобработки и анализа телеметрии открывает путь к проектированию и обучению архитектуры нейронной сети.

Внедрение этих алгоритмов в системы мониторинга создаст фундамент для автоматизации технического обслуживания, что, в свою очередь, минимизирует риски незапланированных простоев серверной инфраструктуры.

Список литературы

[1] Хасти, Т. Элементы статистического обучения: интеллектуальный анализ данных, вывод и прогнозирование / Т. Хасти, Р. Тибширани, Д. Фридман. – 2-е изд. – Москва : Вильямс, 2020. – 768 с.

[2] Кудряшов, Б. Д. Теория информации : учебник для вузов / Б. Д. Кудряшов. – Санкт-Петербург : Питер, 2009. – 320 с.

[3] Герасимов, А. Н. Методы и системы мониторинга технического состояния серверного оборудования / А. Н. Герасимов. – Москва : Техносфера, 2022. – 184 с.

Авторский вклад

Борисенков Матвей Сергеевич – постановка задачи исследования, формирование репрезентативной выборки телеметрических данных, программная реализация алгоритмов фильтрации и масштабирования признаков, проведение рангового корреляционного анализа и расчет метрик информационного прироста.

Шульженко Сергей Николаевич – постановка задачи исследования, концептуальное обоснование применения аппарата теории информации, верификация предложенной методики оптимизации признакового пространства и интерпретация полученных статистических зависимостей.

ATURE SPACE FORMATION FOR PREDICTIVE ANALYSIS OF DISK SUBSYSTEM HEALTH BASED ON S.M.A.R.T. ATTRIBUTES

M.S. Borisenkov
Engineer, PhD Student

S.N. Shuzhenko
Professor of the Department of ITUS A.A. Leonov Technological University (Branch of MIIGAiK), Doctor of Technical Sciences

Abstract. The article proposes a methodology for optimizing the feature space of server telemetry based on correlation analysis and the information entropy apparatus.

Keywords: predictive maintenance, server equipment telemetry, data preprocessing, Spearman correlation coefficient, information gain, Shannon entropy, feature dimensionality reduction.