DOI: 10.18721/JCSTCS.12412
УДК 004.021

ОБЗОР ПОДХОДОВ К ОБНАРУЖЕНИЮ СБОЕВ В СИСТЕМАХ ХРАНЕНИЯ ДАННЫХ

М.Б. Успенский

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

Выполнен обзор существующих программных средств, предназначенных для мониторинга состояния систем хранения данных, определены применяющиеся подходы к сбору, обработке и хранению данных, описаны используемые инструменты для обнаружения сбоев, сформулированы перечни признаков для классификации и сравнения существующих программных решений. На основании проведенного анализа решаемых существующими программными средствами задач предложена типовая архитектура программного комплекса для обнаружения сбоев, описаны входящие в неё модули и характер их взаимодействия. Выполнен обзор актуальных публикаций, посвященных обнаружению сбоев и выявлению аномалий в сфере хранения данных и вычислительных систем, рассмотрены представленные в них алгоритмы, основанные на методах кластеризации и классификации, статистического анализа, опорных векторов, изолирующего леса, искусственных иммунных систем, сетей инвариантов и др.

Ключевые слова: выявление аномалий, машинное обучение, обнаружение сбоев, система хранения данных, локализация неисправностей.

Ссылка при цитировании: Успенский М.Б. Обзор подходов к обнаружению сбоев в системах хранения данных // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2019. Т. 12. № 4. С. 145–158. DOI: 10.18721/JCSTCS.12412

A SURVEY OF THE APPROACHES TO STORAGE SYSTEMS FAULT DETECTION

M.B. Uspenskiy

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

In present paper, we have carried out a comparative analysis of existing software used for health monitoring in enterprise-level storage systems, described commonly used approaches to monitoring data collection, processing and storage, fault detection methods. Based on this analysis we proposed criteria for monitoring software classification and comparison, generalized monitoring software architecture, its modules and module interaction. We also carried out a survey of the recent publications dedicated to anomalies detection, fault diagnosis in a field of data storage and computing systems, and described commonly used algorithms, including clusterization and classification methods, statistical analysis, SVM, isolated forest, artificial immune system, invariant networks.

Keywords: anomaly detection, machine learning, fault diagnosis, storage system, root cause analysis.

Citation: Uspenskiy M.B. A survey of the approaches to storage systems fault detection. St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems, 2019, Vol. 12, No. 4, Pp. 145–158. DOI: 10.18721/JCSTCS.12412

Введение

Системы хранения данных (СХД) корпоративного уровня в настоящее время являются сложными программно-аппаратными продуктами, которые могут включать в себя, кроме собственно носителей информации, одну или несколько управляющих ЭВМ, дисковые шасси, сетевую инфраструктуру, фабрики (одну или несколько), а также системное программное обеспечение, предоставляющее безопасный доступ к данным, организацию объединения физических носителей информации в логические сущности, управление кластерами, репликацию и избыточность данных. Взаимодействие разнородных программных и аппаратных компонентов в условиях высокой нагрузки может приводить к возникновению сбоев в работе СХД даже в том случае, когда неисправности в отдельных компонентах отсутствуют. Например, в [1, 2] указывается, что на сбои в процессе взаимодействия компонентов приходится до 11 % от общего числа сбоев в работе СХД.

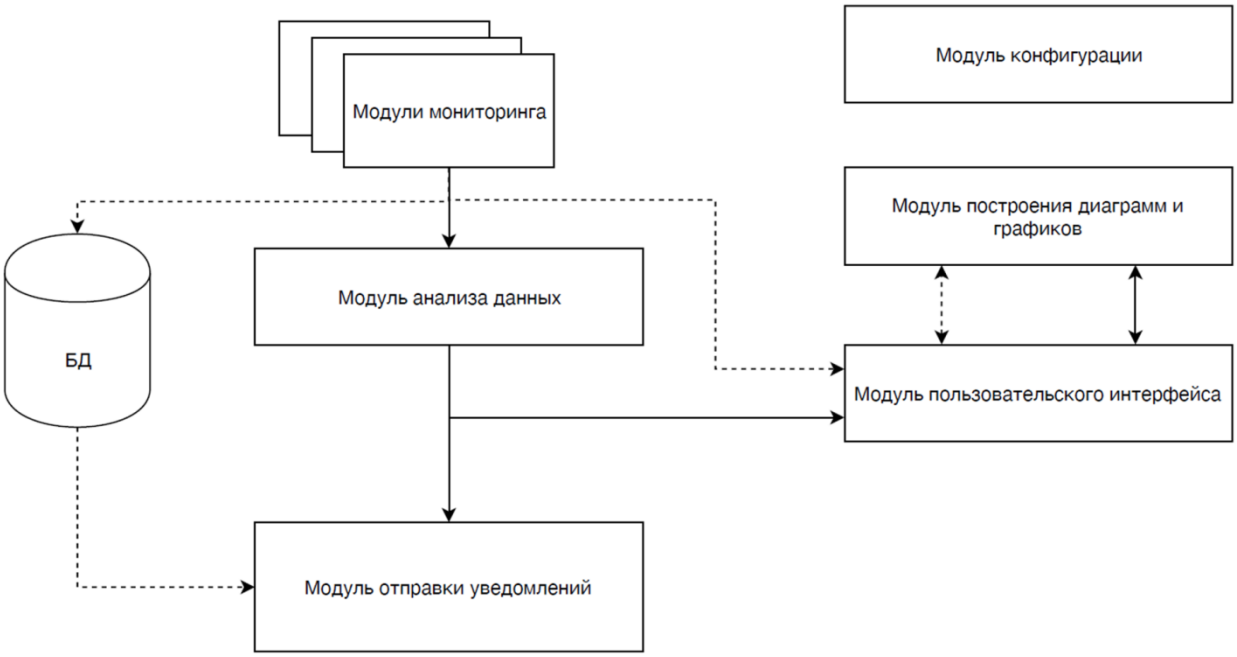
Обнаружение сбоев в процессе функционирования СХД является комплексной задачей, требующей анализа как программных, так и аппаратных компонентов системы, а также процесса их взаимодействия. Актуальность задачи своевременного обнаружения сбоев в работе отдельных программных или аппаратных компонентов СХД и/или СХД в целом определяется тем, что её решение позволяет снизить или устранить вероятность деградации производительности СХД, временной потери доступа к пользовательским данным или потери данных.

Постановка задачи. Цель исследования – доработка программного комплекса прогнозирования сбоев систем хранения данных по результатам исследовательских испытаний. В настоящем обзоре предполагается проанализировать существующие подходы к обнаружению сбоев в области компьютерных систем, применяемые в существующих программных решениях для мониторинга и диагностики систем хранения данных, на основании чего выполнить обзор научных публикаций, предлагающих перспективные методы и средства, которые могут быть применены для эффективной реализации данных подходов. Полученная таким образом информация о достоинствах и недостатках этих методов и средств будет применяться в процессе разработки оптимального подхода к обнаружению сбоев для систем хранения данных.

Обзор существующих программных решений в области обнаружения сбоев в работе СХД

Обобщенная структура программного комплекса, полученная на основе анализа представленных на текущий момент решений в области мониторинга и диагностики, приведена на рисунке.

Архитектура, представленная на рисунке, определяется характером решаемых программным комплексом задач: в простейшем случае (связи обозначены пунктирными линиями) программный комплекс не предназначен для интеллектуального анализа данных мониторинга, и обнаружение сбоев осуществляется путем визуального контроля администратором изменений в наблюдаемой СХД. В более сложных случаях используются встроенные средства анализа результатов мониторинга (связи модулей обозначены простыми линиями). Программные комплексы



Обобщенная структура программного комплекса для обнаружения сбоев в работе СХД
The generalized structure of the software package for detecting failures in the storage system

обнаружения сбоев в работе СХД могут включать в себя все или некоторые из программных инструментов, решающих следующие задачи.

• Сбор метрик производительности и/или здоровья – набор модулей мониторинга. Такие модули могут реализовываться, например, в виде программных агентов (см. Zabbix и др.), размещаемых на вычислительных узлах СХД. Отдельный интерес представляет комплекс VirtualWisdom [3], имеющий в своём составе аппаратные средства мониторинга производительности агентов.

• Анализ метрик производительности и/или здоровья для диагностирования или прогнозирования возникновения сбоев – модуль анализа данных. Анализ собранных данных в зависимости от интеллектуальности системы может осуществляться с использованием широкого спектра методов: от сравнения данных мониторинга с пороговыми значениями до использования алгоритмов машинного обучения. Анализ выполняется с целью определения связей (корреляции) между данными мониторинга и состоянием СХД.

• Информирование управляющего персонала о наступлении тех или иных событий, генерируемых модулями мониторинга или анализа данных – модуль отправки уведомлений. События могут быть как заранее заданные разработчиками программного средства, так и определенные администратором. Для решения этой задачи может использоваться отправка уведомлений по электронной почте, смс и/или мобильным мессенджерам.

• Хранение настроек модулей – модуль конфигурации. Такой модуль должен быть связан со всеми прочими модулями системы (эти связи не представлены на рисунке, чтобы не перегружать его лишними деталями). Как правило, могут храниться правила формирования уведомлений, частота сбора данных мониторинга, перечень данных мониторинга, отображаемых диаграмм и графиков и т. д.

• Управление настройками программного комплекса и визуальный контроль за состоянием СХД – модуль графического интерфейса (например, отображение временных графиков изменения зна-

чений параметров с отображением трендов и уровней среднего значения (ElasticStack)).

• Хранение исторических данных о значениях параметров СХД и выходных данных модуля анализа состояний. Хранение осуществляется с помощью базы данных (в некоторых случаях может быть заменена записями в журналы).

Существующие на сегодняшний день программные средства, применяемые для обнаружения сбоев в работе СХД, можно классифицировать по следующим признакам:

• по способу реализации сбора данных: программные (например, Zabbix, ElasticStack) и программно-аппаратные (VirtualWisdom);

• по характеру анализируемых данных: анализ данных мониторинга (Zabbix), анализ журналов (Splunk), анализ данных в общем виде (ElasticStack, Anomaly.io);

• по области применения: средства, применяемые для СХД определенного производителя (встраиваемые в программные средства управления ресурсами конкретных типов СХД – IBM Spectrum Control, Hitachi System Event Management Tool), диагностические средства общего назначения (Zabbix, ElasticStack), средства анализа данных (Anomaly.io). Диагностические средства общего назначения, как правило, предназначены для мониторинга и диагностики серверных ЭВМ в целом, но используются также в задачах мониторинга и обнаружения сбоев в работе СХД.

В [4] описан процесс разработки программного комплекса обнаружения аномалий в реальном времени Anodot, в том числе предложены следующие критерии, которыми необходимо руководствоваться при проектировании программного комплекса для поиска аномалий:

• время реакции на аномалию (может быть требование реального времени или требование к величине временного интервала, в течение которого необходимо обнаружить аномалию);

• объём метрик, которые надо анализировать;

• частота изменения анализируемых временных рядов;

• связность анализируемых метрик, т. е. необходимо ли оценивать аномалии всех метрик в целом или определять их по каждой конкретной метрике;

• устойчивость при отсутствии внешнего контроля (система может опираться на оценки экспертов, а может предъявляться требование).

Для реализации конкретной системы, реализующей устойчивое обнаружение аномалий в большой системе в режиме реального времени с постоянно меняющимися данными и связными метриками, предлагается следующая последовательность шагов при проектировании и обучении: реализация системы универсального сбора метрик, не требующего ручного конфигурирования, исследование нормального поведения объекта анализа (обучение модели при помощи того, что считается нормальным, и вывод статистического теста для отнесения данных к аномальным, если они не описываются моделью), исследование аномального поведения объекта (разработка классификатора аномалий по типу и значимости), исследование связей между метриками с применением алгоритмов кластеризации, точечное обучение на основании оценки распознанных аномалий как ложноположительных или ложноотрицательных.

В рассмотренных программных комплексах выявлены следующие подходы к обнаружению сбоев.

• Обнаружение сбоев на основании заданного списка событий (например, реализовано в программном обеспечении IBM Spectrum Accelerate, Hitachi System Event Management Tool).

• Обнаружение сбоев на основании заданных пороговых значений. Собираемая метрика сравнивается с пороговыми значениями. В случае выхода значения метрики за пределы пороговых значений формируется уведомление администратору (реализовано, например, в Fujitsu ServerView System Monitor).

• Выявление аномалий. В этом случае собираемые метрики анализируются на предмет наличия редких данных, отличающихся от прочих. Решение задачи выявления аномалий может осуществляться разными способами, в том числе с использованием методов машинного обучения (например, методов теории распознавания образов), статистических методов и т. д. Выявление аномалий является основным способом обнаружения сбоев в большинстве существующих программных средств с использованием интеллектуальных средств анализа данных, в том числе Elastic, Anomaly.io и др.

• Сопоставление топологии системы с получаемыми данными мониторинга с определением перечня компонентов в пути прохождения данных.

• Локализация неисправностей с использованием методик анализа коренных причин (root case analysis) на основании данных мониторинга и выявленных аномалий. Реализовано во многих системах, в автоматическом или автоматизированном режимах, например, Infrastructure Analytics Advisor (Hitachi).

Программные продукты, описанные в данном разделе, приведены в сводной таблице. Учитываются следующие характеризующие их признаки:

тип (анализ данных, система общего назначения, встроенная в продукт конкретного производителя);

обладает ли система открытым исходным кодом;

обладает ли система встроенными средствами сбора данных;

Сводная таблица существующих программных средств

Summary table of existing software

Название	Тип	Открытый код	Средства сбора данных	Наличие средств локализации неисправностей	Наличие встроенных средств анализа собранных данных
Anomaly.io	Анализ данных	–	–	–	Визуализация данных, выявление аномалий
Active Health System (HPE)	Встроенные, HPE	–	+	–	–
Active IQ	Встроенные, NetApp	–	+	+	Пороговые значения для формирования уведомлений, визуализация данных
Cacti	Общего назначения, мониторинг	+	+	–	Визуализация данных
CloudIQ	Встроенные, DELL	–	+	+	Выявление аномалий, визуализация данных, построение трендов, оценка рисков
ElasticStack	Анализ даннных	+	+	–	Выявление аномалий, пороговые значения для формирования уведомлений, визуализация данных
Fujitsu ServerView System Monitor	Встроенные, Fujitsu	–	+	+	–

Окончание таблицы

Название	Тип	Открытый код	Средства сбора данных	Наличие средств локализации неисправностей	Наличие встроенных средств анализа собранных данных
Hitachi System Event Management Tool	Встроенные, Hitachi	–	+	+	Диагностика сбоя по коду события в журнале
Icinga	Общего назначения, мониторинг	–	+	–	Пороговые значения для формирования уведомлений, визуализация данных
IBM Storage Insights	Встроенные, IBM	–	+	+	Пороговые значения для формирования уведомлений, визуализация данных
Nagios	Общего назначения, мониторинг	–	+	–	Пороговые значения для формирования уведомлений, визуализация данных
OnCommand Insight	Встроенные, NetApp	–	+	+	Визуализация данных, пороговые значения для формирования уведомлений, построение трендов
Splunk	Анализ данных	–	+	–	Визуализация данных, пороговые значения для формирования уведомлений
Storage Analytics	Встроенные, Dell EMC	–	+	+	Визуализация данных, пороговые значения для формирования уведомлений
VirtualWisdom	Общего назначения, мониторинг	–	+	+	Визуализация данных, пороговые значения для формирования уведомлений
Zabbix	Общего назначения, мониторинг	+	+	–	Визуализация данных, пороговые значения для формирования уведомлений, выявление аномалий

обладает ли система встроенными средствами локализации неисправностей;

краткое перечисление методов анализа сбора данных (при возможности их идентификации).

Обзор перспективных подходов к решению задачи выявления аномалий в компьютерных системах

Как было показано в предыдущем разделе, наиболее распространенным методом обнаружения сбоев является выявление аномалий. Выявление аномалий обычно ведется в двух направлениях – обнаружение выбросов и обнаружение новизны (в таком случае определяется нормальное поведение системы, и аномалиями считаются объекты, отсутствующие в обучающей выборке). Существуют различные подходы к выявлению аномалий: статистические методы; методы, основанные на кластеризации и классификации; методы с использованием искусственных нейронных сетей. Анализ публикаций, посвященных диагно-

стике сбоев с использованием выявления аномалий за последние пять лет, позволил выделить наиболее перспективные направления в этой области. Далее детально будут рассмотрены следующие группы методов, присутствующие в данных публикациях:

методы, основанные на кластеризации и классификации;

методы, основанные на использовании сети инвариантов;

обнаружение аномалий с применением метода опорных векторов;

искусственные иммунные системы;

прочие методы, в том числе, использующие подходы статистического анализа.

Методы, основанные на кластеризации и классификации. Методы обнаружения аномалий, основанные на решении задач классификации или кластеризации, являются наиболее распространенными в современных работах, посвященных обнаружению сбоев в области диагностики вычислительных устройств или устройств хранения данных.

В [5] предложен подход к диагностике сбоев в реальном времени, основанный на рекурсивной оценке плотности. Оценка плотности опирается на распределение Коши, параметры которого можно обновлять рекурсивно, что позволяет хранить в памяти минимальное количество данных. Преимуществом предложенного подхода является отсутствие требования по наличию модели диагностируемого процесса или исторических данных для формирования обучающей выборки.

В работе [6] изучено использование адаптивной ядерной оценки плотности для определения выбросов в нелинейных системах. В соответствии с этим подходом каждой выборке данных назначается локальное значение выброса, показывающее, насколько одна выборка отличается от других в ее окрестности.

Процесс организации анализа сетевого потока данных на основе кластеризированных образов рассмотрен в [7]. Метод направлен на решение задачи обработки большого объёма исходных данных (десятки миллионов пакетов в секунду на одно сетевое соединение в 10 Гб, что не позволяет прямо применять детальный анализ) путем агрегирования переменных, описывающих трафик в абстрактные образы (для кластеризации в работе используется метод k-средних), после чего задача определения аномалий сводится к задаче сравнения образов.

В работе [8] предложен комбинированный метод диагностики, использующий совместно модельный подход к диагностике и классификаторы аномалий для локализации неисправностей, направленный на совмещение преимуществ обоих подходов и повышение точности реализованного алгоритма путем переобучения классификаторов на основании выявленных в процессе эксплуатации сбоев. Диагностирование происходит по следующей схеме: на основании модельного подхода определяются потенциальные сбои, оценка вероятности которых осуществляется при помощи классификаторов аномалий. С помощью данного подхода решены задачи определения и локализации неисправностей в условиях недостатка данных для обучающих выборок, переобучения классификаторов в автоматическом режиме на основании выявленных сбоев и поиска множественных сбоев в условиях наличия только единичных сбоев в обучающих выборках.

Прикладной случай решения задачи обнаружения сбоев на основании метрик производительности компонентов виртуальных машин с использованием платформы анализа больших наборов данных Hadoop MapReduce совместно с наивным классификатором Байеса в сервисе предоставления виртуальных машин и облачных хранилищ рассмотрен в [9]. Точность полученного метода достигает 89,8 % с ростом объёма поступающих данных.

Масштабируемый непараметрический метод, предназначенный для поиска аномалий производительности в вычислительных системах большого размера, предложен в [10]. Для решения данной задачи представлен децентрализованный подход,

основанный на широко распространенном методе сравнения узлов (описан, например, в [11]), включающий в себя следующие этапы: иерархическая группировка системных узлов, непараметрическая кластеризация и двухступенчатое определение аномальных узлов в каждой группе. Преимущества такого подхода – хорошая масштабируемость и высокая производительность полученного решения.

Сети инвариантов. В последнее время широкое распространение получили методы, основанные на комплексном описании поведения систем с использованием сетей инвариантов. В сетях инвариантов узел представляет собой компонент системы, а дуга – значимое, стабильное взаимодействие между двумя компонентами. Инвариантная модель ориентирована на определение стабильных значимых зависимостей между парами компонентов, которые наблюдаются при помощи записи временных рядов, для определения состояния системы. Сильная связь между компонентами называется *инвариантной* (*корреляционной*) *зависимостью*. Объединяя инварианты, определенные по всем наблюдаемым компонентам, можно получить глобальный профиль системы. В таком случае аномальное поведение системы и источник аномалии определяется на основании поиска поврежденных инвариантов (исчезающие корреляции).

Подход к определению неисправностей в программном обеспечении, использующий рассогласование динамических инвариантов, предложен в [12]. Подход направлен на решение таких проблем динамических инвариантов, как высокая вычислительная сложность определения инвариантов, фильтрации ложноположительных срабатываний и устранение избыточности. Смысл подхода заключается в предварительной фильтрации подозрительных функций и последовательном применении к ним инструментов определения инвариантов.

Рассмотренный в [13] подход по определению причинно-следственных аномалий при помощи сетей инвариантов с использованием временного и динамического анализа исчезающих корреляций направлен на решение таких проблем метода сетей инвариантов, как невозможность определения маршрута распространения сбоя по сети, наличия узлов с максимальным процентом исчезающих корреляций, не являющихся при этом корневыми случайными аномалиями, отсутствие возможности использовать априорные знания об аномальных узлах и временных параметрах исчезающих корреляций. В качестве инструментов для решения этих проблем предлагается сетевая диффузия для моделирования распространения причинно-следственных аномалий, применение оценок маршрута распространяющихся причинно-следственных аномалий для реконструкции исчезающих корреляций, а также использование стратегии нормализации, позволяющей не учитывать экстремальные значения или выбросы без необходимости явно исключать их из набора данных. Кроме того, алгоритм позволяет использовать априорное знание для оценки состояния отдельных аномальных узлов в некоторые моменты времени, с поправкой на то, что такое априорное знание может быть сильно зашумленным.

Метод опорных векторов. Давно использующийся для задач классификации метод опорных векторов также часто применяется для выявления аномалий.

В исследовании [14] описан случай применения метода опорных векторов для определения аномальных виртуальных машин в облачной среде с целью предотвращения деградации производительности. Для этих целей авторами разработана программная платформа EaAD, позволяющая определять аномалии в виртуальных машинах с учетом влияния внешней среды. Для оценки виртуальных машин используются их метрики производительности. Для определения аномалий в EaAD реализуются алгоритмы, основанные на различных вариантах метода опорных векторов.

Модификация метода опорных векторов для одного класса, направленная на решение проблемы уязвимости метода к выбросам в обучающей выборке, предложена в [15]. Разработанный устойчивый метод опорных векторов для одного класса подразумевает подавление возникающих выбросов в обучающей выборке за счет введения отрицательных весовых коэффициентов.

Искусственные иммунные системы. Искусственные иммунные системы – это вычислительные системы, основанные на принципах работы иммунной системы. Они включают в себя в том числе негативный алгоритм отбора, иммунный сетевой алгоритм и пр. Использование иммунных систем в качестве основы для алгоритмов выявления аномалий обосновывается тем, что биологические иммунные системы, так же как и алгоритмы выявления аномалий, нацелены на выявление информации, отличной от нормальной [16, 17].

В работе [18] предложено использовать для обнаружения сбоев две разновидности алгоритма негативного отбора – FB-NSA (Fixed Boundary NSA) и его доработанный вариант FFB-NSA (Fine Fixed Boundary NSA). Принципиальное отличие от прочих работ в области алгоритмов отрицательного отбора (например, в [19] предложен детектор в виде гиперкуба, в [20] – детектор в виде гиперэллипсоида, в [21] – детектор с множественной формой) заключается в направленности не на исследование формы детекторов, а на то, чтобы добиться их неизменности. Константный детектор зависит только от обучающей выборки и не связан с временем обучения. Оба предложенных алгоритма генерируют слой детекторов, расположенных между нормальными и аномальными обучающими данными. При этом детекторы имеют постоянное число, размер и положение. Доработанный алгоритм заключает в себе механизм охвата слепой зоны, возникающей при использовании больших детекторов и с малым значением параметра m (размера тестовой выборки).

Прочие методы. В работе [22] предложено выполнять обнаружение сбоев и определение источника их возникновения на основе графической вероятностной модели. Метод предлагает альтернативу использования традиционной Байесовой сети в качестве графической модели в виде модели причинно-следственных связей системы, с применением ядерной оценки плотности для оценки вероятностной функции плотности вместо изучения параметров, характерного для Байесовой сети.

Подход к диагностике жестких дисков на основании полупараметрических моделей и статистических оценок рассмотрен в [23]. Для моделирования распределения параметров SMART исправного жесткого диска используется смесь Гауссовских распределений. Аномалиями при этом считаются случаи, когда рассчитанная статистическая оценка различий между поведением диска и полученной модели превышает пороговое значение. В работе [24] рассмотрено сравнение алгоритма, основанного на смеси Гауссовских распределений с несколькими алгоритмами, в том числе, алгоритмом, основанным на использовании расстояния Махалнобиса, алгоритмом, основанным на методе опорных векторов, и стандартным алгоритмом диагностики SMART, и демонстрируется его преимущество в точности. Для оценки качества работы алгоритмов используются метрики FDR (fault detection rate) и FAR (false alarm rate) на реальном наборе статистических данных о сбоях жестких дисков. Алгоритм, основанный на смеси Гауссовских распределений, оказывается наиболее эффективным, с FDR 80,59 % и с FAR 0 %;

Алгоритм, основанный на латентной корреляционной вероятностной модели, предназначенный для поиска аномалий в данных мониторинга состояния оборудования, представлен в [25]. Предложенный алгоритм направлен на решение основных проблем, возникающих в процессе поиска аномалий в данных мониторинга: необходимости обработки большого объёма дан-

ных, генерируемых системами мониторинга, сложности идентификации источника аномалии (т. к. в общем случае аномалия в одном компоненте оборудования влияет на данные мониторинга сразу нескольких связанных с ним компонентов) и наличия шума в данных мониторинга. Для решения этих проблем вводится понятие латентной корреляции, которая определяет связь между различными наборами данных мониторинга в некоторый момент времени.

Подход к обнаружению сбоев для одного класса, основанный на использовании генеративно-состязательных сетей, предложен в [26]. Генеративно-состязательная сеть является алгоритмом машинного обучения без учителя, в котором одна нейронная сеть генерирует обучающие выборки для другой нейронной сети, пытающейся выявить в созданных выборках аномалии (см., например, описание генеративной сети в [27]). В подходе, описанном в [26], первая нейронная сеть пытается определить нормальное протекание процесса, а вторая принимает окончательное решение о наличии аномалий. Для обеспечения конвергентности системы в процессе обучения и для повышения точности классификатора предложено использовать специализированный алгоритм, проверяющий обученные модели на специальном валидационном наборе данных. Полученный подход позволяет реализовать алгоритм более производительный, чем классические методы опорных векторов и изолирующего леса.

Прикладная задача выявления ранее не выявленных аномалий в полётных данных самолетов, собранных в единой базе данных, рассмотрена в [28]. Для решения данной задачи предложено использовать комбинированный подход с применением методов формализации экспертного знания и методов машинного обучения без учителя. Полученный в итоге алгоритм имеет неспециализированный характер и применим к выявлению аномалий в различных промышленных системах. Для решения проблемы большого объёма анализируемых данных используется вейвлет-преобразование, позволяющее преобразовать непрерывный временной ряд в набор дискретных отсчетов, сохраняющих пространственно-временные характеристики сигнала, после чего этот набор сворачивается в матрицу сходства, к которой затем применяется алгоритм иерархической кластеризации, позволяющий построить набор кластеров, каждый из которых содержит схожие по параметрам сведения о полетах. Далее определяются параметры, позволяющие получить максимальное расхождение между кластерами, после чего данные анализируются экспертами и применяются для обучения алгоритмов с учителем.

Обнаружение аномалий в условиях смешанных численно-категориальных исходных данных большого масштаба изучено в [29]. Для решения этой задачи предложена комбинированная вероятностная мера, представляющая собой маргинальную плотность для непрерывных переменных и условную вероятность для категориальных переменных. В результате итоговый алгоритм, базирующийся на модели, обученной с помощью целевой функции максимального сходства и оптимизированной с помощью стохастического градиентного спуска, демонстрирует наиболее высокую точность среди всех алгоритмов определения аномалий по нескольким наборам параметров, а также максимальную производительность и масштабируемость.

Заключение

В ходе анализа актуальных тенденций в области диагностики систем хранения данных корпоративного уровня рассмотрены наиболее распространенные программные системы управления и мониторинга СХД, анализа данных и программные средства общего назначения. На текущий момент можно сделать вывод о том, что почти все рассмотренные актуальные программные средства предлагают некоторую функциональность для автоматического, автоматизированного или выполняемого администратором поиска неисправностей варьируемой сложности [30, 31].

Кроме того, можно сделать вывод о том, что системы общего назначения, предоставляющие унифицированную функциональность по сбору, обработке и визуализации информации, получают всё большее распространение. Преимуществом таких систем является использование ими большого количества разнообразных алгоритмов выявления аномалий, в том числе вынесенных в сферу облачного вычисления, что позволяет экономить вычислительные ресурсы СХД. При этом алгоритмы выявления аномалий сами по себе не обладают знаниями о топологии и устройстве конкретных систем хранения данных и, следовательно, в общем случае не подходят для определения типов возникающих неисправностей, пути распространения ошибки по топологии системы и конкретных компонентов, в которых возникла неисправность. Программные средства, предоставляемые производителями систем хранения данных, имеют более слабое алгоритмическое обеспечение по выявлению аномалий, но за счет наличия у разработчиков экспертного знания о предмете анализа, обладают, как правило, инструментами локализации неисправности.

Анализ публикаций, посвященных обнаружению неисправностей или выявлению аномалий, позволяет сделать вывод о том, что в настоящее время широкое распространение получают алгоритмы, комбинирующие различные группы методов выявления аномалий. Задача повышения точности алгоритмов во многих случаях отходит на второй план, делая более приоритетными задачи повышения производительности, экономии системных ресурсов, масштабируемости и т. д. В рассмотренных публикациях применяются как классические группы методов, такие как методы на основе опорных векторов, методы, основанные на алгоритмах кластеризации и классификации и т. д., так и более редкие, представляющие, возможно, больший интерес, такие как алгоритмы искусственных иммунных сетей.

Для разработки подходов к обнаружению и локализации неисправностей в работе СХД перспективной представляется такая организация системы обнаружения сбоев, при которой возможно комбинировать формализованное экспертное знание о предметной области с получаемой извне информацией о наличии аномалий в данных мониторинга, т. е., фактически, параллельное применение методов, основанных на модельно-ориентированном подходе и подходе, ориентированном на анализе массива данных.

Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы». Уникальный идентификатор RFMEFI58117X0023.

СПИСОК ЛИТЕРАТУРЫ

1. Hewlett Packard Enterprise Development LP. Can Machine Learning Prevent Application Downtime? // URL: https://cdm-cdn.nimblestorage.com/2017/08/23090218/Can-Machine-Learning-White-Paper-1708-FINAL-print.pdf (Дата обращения: 04.04.2019).

2. **Wang D.** Artificial intelligence makes flash storage predictive // URL: https://www.hpe.com/us/en/insights/articles/artificial-intelligence-makes-flash-storage-predictive-1803.html (Дата обращения: 01.04.2019)

3. **Lelii S.** VirtualWisdom adds storage probe for NetApp NAS array // URL: https://searchstorage.techtarget.com/news/2240219781/VirtualWisdom-adds-storage-probe-for-NetApp-NAS-array (Дата обращения: 20.04.2019).

4. **Toledano M., Cohen I., Ben-Simhon Y., Tadeski I.** Real-time anomaly detection system for time series at scale // Proc. of the KDD 2017: Workshop on Anomaly Detection in Finance. 2018. Pp. 56–65.

5. **Costa B.S.J., Angelov P.P., Guedes L.A.** Real-time fault detection using recursive density estimation // J. of Control, Automation and Electrical Systems. 2014. Vol. 4. No. 25. Pp. 428–437. DOI: 10.1007/s40313-014-0128-4

6. **Zhang L., Lin J., Karim R.** Adaptive kernel density-based anomaly detection for nonlinear systems // Knowledge-Based Systems. 2018. No. 139. Pp. 50–63. DOI: 10.1016/j.knosys.2017.10.009

7. **Kim J., Sim A., Tierney B., Suh S., Kim I.** Multivariate network traffic analysis using clustered

patterns // Computing. 2018. Vol. 101. No. 4. Pp. 339–361. DOI:10.1007/s00607-018-0619-4

8. **Jung D., Ng K.Y., Frisk E., Krysander M.** Combining model-based diagnosis and data-driven anomaly classifiers for fault isolation // Control Engineering Practice. 2018. No. 80. Pp. 146–156. DOI: 10.1016/j.conengprac.2018.08.013

9. **Alkasem A., Liu H., Shafiq M.** Improving fault diagnosis performance using Hadoop MapReduce for efficient classification and analysis of large data sets // J. of Computers. 2018. Vol. 29. No. 4. Pp. 185–202.

10. **Yu L., Lan Z.** A scalable, non-parametric method for detecting performance anomaly in large scale computing // IEEE Transactions on Parallel and Distributed Systems. 2016. Vol. 27. No. 7. Pp. 1902–1914.

11. **Lan Z., Zheng Z., Li Y.** Toward automated anomaly identification in large-scale systems // IEEE Trans. Parallel Distrib. Syst. 2010. Vol. 21. Pp. 174–187.

12. **Wang X., Liu Y.** Fault localization using disparities of dynamic invariants // J. of Systems and Software. 2016. Vol. 122. Pp. 144–154. DOI: 10.1016/j.jss.2016.09.014

13. **Cheng W., Ni J., Zhang K., Chen H., Jiang G., Shi Y., Wang W.** Ranking causal anomalies for system fault diagnosis via temporal and dynamical analysis on vanishing correlations // ACM Transactions on Knowledge Discovery from Data. 2017. Vol. 11. No. 4. Pp. 1–28. DOI:10.1145/3046946

14. **Wang G.P., Wang J.W.** An anomaly detection framework for detecting anomalous virtual machines under cloud computing environment // Internat. J. of Security and Its Applications. 2016. Vol. 10. No. 1. Pp. 75–86.

15. **Yin S., Zhu X., Jing C.** Fault detection based on a robust one class support vector machine // Neurocomputing. 2014. Vol. 145. Pp. 263–268. DOI: 10.1016/j.neucom.2014.05.035

16. **Dasgupta D., González F.** An immunity-based technique to characterize intrusions in computer networks // IEEE Trans. Evol. Comput. 2002. Vol. 6. Pp. 281–291.

17. **Silva G.C., Caminhas W.M., Palhares R.M.** Artificial immune systems applied to fault detection and isolation: A brief review of immune response-based approaches and a case study // Applied Soft Computing. 2017. Vol. 57. Pp. 118–131.

18. **Li D., Liu S., Zhang H.** Negative selection algorithm with constant detectors for anomaly detection // Applied Soft Computing. 2015. Vol. 36. Pp. 618–632.

19. **González F., Gymez J., Dasgupta D.** An evolutionary approach to generate fuzzy anomaly (attack) signatures // Proc. of IEEE Systems, Man and Cybernetics Society. IEEE, 2003. Pp. 251–259.

20. **Shapiro J.M., Lamont G.B., Peterson G.L.** An evolutionary algorithm to generate ellipsoid network intrusion detectors // Proc. of the 2005 Workshops on Genetic and Evolutionary Computation. ACM, 2005. Pp. 178–180.

21. **Balachandran S., Dasgupta D., Nino F., Garrett D.** A framework for evolving multi-shaped detectors in negative selection // Proc. of IEEE Symp. on Foundations of Computational Intelligence. IEEE, 2007. Pp. 401–408.

22. **Chen X., Wang J., Zhou J.** Fault detection and backtrace based on graphical probability Model // Prognostics and System Health Management Conf. (PHM-Chongqing). IEEE, 2018. Pp. 584-590.

23. **Queiroz L.P., et al.** Fault detection in hard disk drives based on a semi parametric model and statistical estimators // New Generation Computing. 2018. Vol. 36. No. 1. Pp. 5–19.

24. **Queiroz L.P., et al.** Fault detection in hard disk drives based on mixture of Gaussians // 5th Brazilian Conf. on Intelligent Systems. IEEE, 2016. Pp. 145–150.

25. **Ding J., et al.** An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model // Applied Intelligence. 2016. Vol. 44. No. 2. Pp. 340–361.

26. **Plakias S., Boutalis Y.S.** Exploiting the generative adversarial framework for one-class multi-dimensional fault detection // Neurocomputing. 2019. Vol. 332. Pp. 396–405.

27. **Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.** Generative Adversarial Networks. 2016. arXiv:1406.2661[stat.ML]

28. **Mack D.L.C., Biswas G., Khorasgani H., Mylaraswamy D., Bharadwaj R.** Combining expert knowledge and unsupervised learning techniques for anomaly detection in aircraft flight data // Automatisierungstechnik. 2018. Vol. 66. No. 4. Pp. 291–307. DOI:10.1515/auto-2017-0120

29. **Eiras-Franco C., et al.** Large scale anomaly detection in mixed numerical and categorical input spaces // Information Sciences. 2019. Vol. 487. Pp. 115–127.

30. **Higley L.** Storage analytics: Can we put any more lipstick on that pig? // URL: https://cloud.kapostcontent.net/pub/3da21605-fc17-4712-991a-1c49dc77b871/mfx131e-pc-mon-130-higleyl.pdf?kui=xxPHjAO870Nzv0HTEjftEw (Дата обращения: 10.04.2019).

31. **Gopisetty S.** Evolution of storage management: Transforming raw data into information // IBM Journal of Research and Development. 2008. Vol. 52. No 4.5. Pp. 341–352.

Статья поступила в редакцию 10.05.2019.

REFERENCES

1. Hewlett Packard Enterprise Development LP. Can Machine Learning Prevent Application Downtime? Available: *https://cdm-cdn.nimblestorage.com/2017/08/23090218/Can-Machine-Learning-White-Paper-1708-FINAL-print.pdf* (Accessed: 04.04.2019).

2. **Wang D.** Artificial intelligence makes flash storage predictive. Available: *https://www.hpe.com/us/en/insights/articles/artificial-intelligence-makes-flash-storage-predictive-1803.html* (Accessed: 01.04.2019).

3. **Lelii S.** VirtualWisdom adds storage probe for NetApp NAS array. Available: *https://searchstorage.techtarget.com/news/2240219781/ VirtualWisdom-adds-storage-probe-for-NetApp-NAS-array* (Accessed: 20.04.2019).

4. **Toledano M., Cohen I., Ben-Simhon Y., Tadeski I.** Real-time anomaly detection system for time series at scale. *Proceedings of the KDD 2017: Workshop on Anomaly Detection in Finance*, 2018, Pp. 56–65.

5. **Costa B.S.J., Angelov P.P., Guedes L.A.** Real-Time Fault Detection Using Recursive Density Estimation. *Journal of Control, Automation and Electrical Systems*, 2014, Vol. 4, No. 25, Pp. 428–437. DOI:10.1007/s40313-014-0128-4

6. **Zhang L., Lin J., Karim R.** Adaptive kernel density-based anomaly detection for nonlinear systems. *Knowledge-Based Systems*, 2018, No. 139, Pp. 50–63. DOI: 10.1016/j.knosys.2017.10.009

7. **Kim J., Sim A., Tierney B., Suh S., Kim I.** Multivariate network traffic analysis using clustered patterns. *Computing*, 2018, Vol. 101, No. 4, Pp. 339–361. DOI:10.1007/s00607-018-0619-4

8. **Jung D., Ng K.Y., Frisk E., Krysander M.** Combining model-based diagnosis and data-driven anomaly classifiers for fault isolation. *Control Engineering Practice*, 2018, No. 80, Pp. 146–156. DOI: 10.1016/j.conengprac.2018.08.013

9. **Alkasem A., Liu H., Shafiq M.** Improving fault diagnosis performance using Hadoop MapReduce for efficient classification and analysis of large data sets. *Journal of Computers*, 2018, Vol. 29, No. 4, Pp. 185–202.

10. **Yu L., Lan Z.** A scalable, non-parametric method for detecting performance anomaly in large scale computing. *IEEE Transactions on Parallel and Distributed Systems*, 2016, Vol. 27, No. 7, Pp. 1902–1914.

11. **Lan Z., Zheng Z., Li Y.** Toward automated anomaly identification in large-scale systems. *IEEE Trans. Parallel Distrib. Syst.*, 2010, Vol. 21, Pp. 174–187.

12. **Wang X., Liu Y.** Fault localization using disparities of dynamic invariants. *Journal of Systems and Software*, 2016, Vol. 122, Pp. 144–154. DOI: 10.1016/j.jss.2016.09.014

13. **Cheng W., Ni J., Zhang K., Chen H., Jiang G., Shi Y., Wang W.** Ranking causal anomalies for system fault diagnosis via temporal and dynamical analysis on vanishing correlations. *ACM Transactions on Knowledge Discovery from Data*, 2017, Vol. 11, No. 4, Pp. 1–28. DOI:10.1145/3046946

14. **Wang G.P., Wang J.W.** An anomaly detection framework for detecting anomalous virtual machines under cloud computing environment. *International Journal of Security and its Applications*, 2016, Vol. 10, No. 1, Pp. 75–86.

15. **Yin S., Zhu X., Jing C.** Fault detection based on a robust one class support vector machine. *Neurocomputing*, 2014, Vol. 145, Pp. 263–268. DOI: 10.1016/j.neucom.2014.05.035

16. **Dasgupta D., González F.** An immunity-based technique to characterize intrusions in computer networks. *IEEE Trans. Evol. Comput.*, 2002, Vol. 6, Pp. 281–291.

17. **Silva G.C., Caminhas W.M., Palhares R.M.** Artificial immune systems applied to fault detection and isolation: A brief review of immune response-based approaches and a case study. *Applied Soft Computing*, 2017, Vol. 57, Pp. 118–131.

18. **Li D., Liu S., Zhang H.** Negative selection algorithm with constant detectors for anomaly detection. *Applied Soft Computing*, 2015, Vol. 36, Pp. 618–632.

19. **González F., Gymez J., Dasgupta D.** An evolutionary approach to generate fuzzy anomaly (attack) signatures. *Proceedings of IEEE Systems, Man and Cybernetics Society*, IEEE, 2003, Pp. 251–259.

20. **Shapiro J.M., Lamont G.B., Peterson G.L.** An evolutionary algorithm to generate ellipsoid network intrusion detectors. *Proceedings of the 2005 Workshops on Genetic and Evolutionary Computation*, ACM, 2005, Pp. 178–180.

21. **Balachandran S., Dasgupta D., Nino F., Garrett D.** A framework for evolving multi-shaped detectors in negative selection. *Proceedings of IEEE Symposium on Foundations of Computational Intelligence*, IEEE, 2007, Pp. 401–408.

22. **Chen X., Wang J., Zhou J.** Fault detection and backtrace based on graphical probability model. *Prognostics and System Health Management Conference (PHM-Chongqing)*, IEEE, 2018, Pp. 584–590.

23. **Queiroz L.P., et al.** Fault detection in hard disk drives based on a semi parametric model and statistical estimators. *New Generation Computing*, 2018, Vol. 36, No. 1, Pp. 5–19.

24. **Queiroz L.P., et al.** Fault detection in hard disk drives based on mixture of Gaussians. *5th Brazilian Conference on Intelligent Systems*, IEEE, 2016, Pp. 145–150.

25. **Ding J., et al.** An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model, *Applied Intelligence*, 2016, Vol. 44, No. 2, Pp. 340–361.

26. **Plakias S., Boutalis Y.S.** Exploiting the generative adversarial framework for one-class multi-dimensional fault detection. *Neurocomputing*, 2019, Vol. 332, Pp. 396–405.

27. **Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.** *Generative Adversarial Networks.* 2016. arXiv:1406.2661[stat.ML]

28. **Mack D.L.C., Biswas G., Khorasgani H., Mylaraswamy D., Bharadwaj R.** Combining expert knowledge and unsupervised learning techniques for anomaly detection in aircraft flight data. *Automatisierungstechnik*, 2018, Vol. 66, No. 4, Pp. 291–307. DOI:10.1515/auto-2017-0120

29. **Eiras-Franco C., et al.** Large scale anomaly detection in mixed numerical and categorical input spaces. *Information Sciences*, 2019, Vol. 487, Pp. 115–127.

30. **Higley L.** Storage analytics: Can we put any more lipstick on that pig? Available: *https://cloud.kapostcontent.net/pub/3da21605-fc17-4712-991a-1c49dc77b871/mfx131e-pc-mon-130-higleyl.pdf?kui=xxPHjAO870Nzv0HTEjftEw* (Accessed: 10.04.2019).

31. **Gopisetty S.** Evolution of storage management: Transforming raw data into information. *IBM Journal of Research and Development*, 2008, Vol. 52, No 4.5, Pp. 341–352.

Received 10.05.2019.

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

УСПЕНСКИЙ Михаил Борисович
USPENSKIY Mikhail B.
E-mail: umihail@list.ru