МОНИТОРИНГ И ДИАГНОСТИКА АНОМАЛЬНЫХ СОСТОЯНИЙ КОМПЬЮТЕРНОЙ СЕТИ НА ОСНОВЕ ИЗУЧЕНИЯ "ИСТОРИЧЕСКИХ ДАННЫХ"

Шелухин Олег Иванович,
Московский технический университет связи и информатики, Москва, Россия, ***sheluhin@mail.ru***

Осин Андрей Владимирович,
Московский технический университет связи и информатики, Москва, Россия, ***osin_a_v@mail.ru***

Костин Денис Владимирович,
Московский технический университет связи и информатики, Москва, Россия, ***d.v.kostin@mail.ru***

DOI: 10.36724/2072-8735-2020-14-4-23-30

Ключевые слова: *аномальные состояния компьютерной сети, прогнозирование, машинное обучение, интеллектуального анализа данных, мониторинг системных показателей, кластеризация, секвенциальный анализ, паттерн*

Предложено характеризовать "здоровье компьютерной сети (КС)" рядом системных показателей, характеризующих уровень обслуживания (Service Level Objectives, SLO), и соглашение об уровне предоставляемого сервиса SLA (Service Level Agreement) рассматриваемой компьютерной сети. Для выполнения поставленных целей можно извлекать из собранных исторических данных необходимые параметры (атрибуты, сигнатуры), определяющие состояние КС, вызванные проблемами, которые могут быть использованы для автоматической кластеризации и, на основе сходства, поиска подобных проблем "в прошлом". Нахождение таких событий в базе данных позволяет сопоставить наблюдаемое текущее поведение системы с ранее возникшей проблемой и определить, является ли причина текущей проблемы аналогичной, уже имевшей место в прошлом. Для решения подобной задачи необходимо на этапе обучения изучить различные аномальные симптомы из исторических данных. Для прогнозирования "будущих" симптомов, необходимо смоделировать статистические изменения шаблонов (паттернов) различных значений признаков. Предложена функциональная схема системы диагностики здоровья, а также предсказания рисков ухудшения "здоровья" компьютерной сети. Анализируются характеристики используемые в процессе диагностики здоровья компьютерной сети. Сочетая классификацию симптомов аномалии и предсказание, система диагностики должна выполнять прогнозирование аномалий в сети, то есть выполнять классификацию симптомов аномалии для будущих данных. Предложенное алгоритмическое и программное решение может быть использовано в системах мониторинга качества функционирования компьютерных систем, диагностики возникающих проблем, а также раннего обнаружения (на основе алгоритмов предсказания) появления рисков снижения качества функционирования компьютерной сети.

Информация об авторах:
Шелухин Олег Иванович, *д.т.н., МТУСИ, Москва, Россия*
Осин Андрей Владимирович, *к.т.н., МТУСИ, Москва, Россия*
Костин Денис Владимирович, *аспирант МТУСИ, Москва, Россия*

Для цитирования:
Шелухин О.И., Осин А.В., Костин Д.В. Мониторинг и диагностика аномальных состояний компьютерной сети на основе изучения "исторических данных" // T-Comm: Телекоммуникации и транспорт. 2020. Том 14. №4. С. 23-30.

For citation:
Sheluhin O.I., Osin A.V., Kostin D.V. (2020) Monitoring and diagnostics of anomalous states in a computer network based on the study of "historical data". *T-Comm*, vol. 14, no.4, pp. 23-30. (*in Russian*)

Введение

Под мониторингом аномальных состояний компьютерной сети (КС) понимают постоянное наблюдение за КС в поисках аномальных состояний или неисправностей. На этапе мониторинга выполняется сбор первичных данных о работе КС: статистики о количестве циркулирующих в сети кадров и пакетов различных протоколов, состоянии портов концентраторов, коммутаторов и маршрутизаторов и т.п. На этапе последующей обработки осуществляется интеллектуальный анализ собранной на этапе мониторинга информации, сопоставление ее с данными, полученными ранее, и выработки предположений о возможных причинах ненадежной работы КС. Задачи мониторинга решаются программными и аппаратными измерителями, тестерами, сетевыми анализаторами, встроенными средствами мониторинга коммуникационных устройств, а также агентами систем управления. Задача анализа требует более активного участия человека и использования таких сложных средств, как экспертные системы.

Под диагностикой сети принято понимать измерение характеристик работы сети в процессе ее эксплуатации. На сегодняшний день нет достоверного способа диагностирования проблем, возникающих в процессе функционирования КС на основе анализа и обработки событий, происходивших и наблюдаемых в прошлом. В результате, если анализируемые проблемы были ранее решены, можно обосновать их диагноз в настоящем и, возможно, применить эти действия повторно. Даже если проблема осталась нерешенной, можем собрать статистику частоты возникновения подобных повторных проблем, накапливая необходимую информацию для уточнения текущей диагностики наблюдаемых событий.

Для выполнения поставленных целей можно извлекать из системы описание в виде некоторых параметров (атрибутов, сигнатур), которые определяет состояние системы, вызванное проблемами, и могут быть использованы для автоматической кластеризации и поиска подобных проблем в прошлом на основе сходства. Нахождение таких событий в базе данных позволяет сопоставить наблюдаемое текущее поведение системы с ранее возникшей проблемой и определить, является ли причина текущей проблемы аналогичной, уже имевшей место в прошлом.

Обзор работ по теме исследования

Прогнозирование сбоев в режиме онлайн является недостаточно изученной темой, хотя и существует несколько методов, которые используют для различных типов КС [1,2,3]. Так в работе [1] изучалась возможность прогнозирования сбоев компьютерных систем в режиме реального времени с использованием полумарковских цепей в сочетании с кластеризацией. Работа состояла в том, чтобы сравнить предлагаемый метод прогнозирования сбоев под названием «Предсказание похожих событий», с двумя другими хорошо известными методами прогнозирования. Их результаты показали многообещающее улучшение эффективности прогнозов по сравнению с другими популярными методами прогнозирования с точностью до 80%. Однако предложенная модель основывалась на наличии полной априорной информации о моделях прогнозируемых ошибок. При возникновении новой, неизвестной формы ошибки, предложенная модель не могла ее предсказать.

В работе [4] изучалось прогнозирование перегрузки сети в беспроводных сетях путем анализа флагов TCP Ack. Для выявления ненормальных уровней потери пакетов в беспроводной сети использовался метод многошаговой кластеризации. Результаты также выглядели многообещающими и показали высокую эффективность в определении потери пакетов. Однако метод ограничивается анализом только пакетов TCP и UDP.

В работе [5] представлен метод для автоматического извлечения сигнатур из функционирующей системы, которые определяют существенные характеристики состояния и могут быть использованы для автоматической кластеризации и извлечения похожих событий для идентификации состоящий наблюдаемой системы в прошлом. Это позволяет операторам идентифицировать и измерить частоту повторяющихся проблем на различных установках в одних и тех же местах. Показано, что наивный подход, основанный на простой записи «сырых» показателей системы, неэффективен и предложен более сложный подход, основанный на статистическом моделировании. Предложенный метод требует только системных метрик (например, среднее время ответа на транзакцию) операционной системы, которые могут быть успешно собраны существующими коммерческими решениями для мониторинга. Даже если собранные данные не имеют размеченных периодов наблюдаемых инцидентов (что типично), подобный подход успешно кластеризует состояния системы в соответствующие похожие проблемы, позволяя диагностировать и идентифицировать повторные проблемы и характеризовать «синдром» группы проблем.

Опираясь на методы классификации шаблонов и вероятностное моделирование, алгоритмы в работах [6, 7] в качестве выходных значений формируются ансамбль вероятностных моделей, характеризующих состояние в зависимости от некоторых системных показателей. Каждая из этих моделей по сути представляет взаимосвязь показателей системы и состояния системных показателей, характеризующих уровень обслуживания (Service Level Objectives, SLO). Для определения вероятности таких состояний используются древовидные байесовские модели. Из полученных распределений можно извлечь искомую характеристику и оценить её влияние на состояние SLO.

В работе [8] предложена схема для преждевременного оповещения о предстоящих аномалиях системы и оценки возможных причин аномалии. С этой целью для выявления различных симптомов и причин аномалий использованы байесовские методы классификации. Для построения изменяющихся моделей различных показателей измерения вводятся Марковские модели. В результате для предсказания появления аномалии системы будущем и возможных причин возникновения аномалии предложена схема, объединяющая марковские модели и байесовские методы классификации.

Во всех известных работах, посвященных мониторингу и прогнозированию аномальных состояний *компьютерной сети* одной из важнейших задач анализа данных является выделение закономерностей. Для последовательных данных эту задачу решает в частности такая область интеллектуального анализа данных (data mining) как секвенциальный анализ или анализ последовательностных паттернов (sequential pattern mining), позволяющий выявлять часто встречающиеся участки в последовательностях наборов элементов, ана-

лизируемых данных. Секвенциальный анализ (sequential pattern mining, поиск/добыча последовательностных шаблонов) – это разновидность интеллектуального анализа данных (data mining) [9, 10, 11]. Объектом секвенциального анализа является база последовательностей. Целью секвенциального анализа является получение часто встречающихся подпоследовательностей, которые называются последовательностными шаблонами, или последовательностными паттернами [12, 13, 14].

Последовательный анализ паттернов состоит из обнаружении интересных подпоследовательностей в наборе последовательностей, где значимость той или иной подпоследовательности может быть измерена с точки зрения различных критериев, таких как частота появления, длина и др. Последовательный анализ шаблонов имеет множество реальных применений, поскольку данные кодируются в виде последовательностей во многих областях, таких как биоинформатика, электронное обучение, анализ корзины, анализ текста и анализ кликов на веб-страницах [15, 16, 17].

Постановка задачи

Значительный практический интерес представляет создание классификатора проблемных мест КС, который включает в себя множество показателей, которые могут коллективно фиксировать отличительные признаки различных таких проблемных мест. Если в процессе наблюдения за работающей КС «в прошлом» накоплено достаточное количество помеченных данных в выбранном пространстве признаков, то представляется возможным построить изучить модель поведения для классификации немаркированных точек в пространстве функций «в будущем». Для того, чтобы предвидеть проблемные места, необходимо применить классификатор к «будущим» данным. Таким образом, важной задачей является предсказание будущих данных в пространстве признаков на один, два и более временных шага [18].

Один из возможных результатов подобной задачи предсказания состоит в том, что прогнозируемая точка поведения КС на n-м шаге измерения попадает в кластер, представляющий I-й аномальный симптом. Если этот результат имеет большую вероятность, система должна предупредить, что аномалия с I-м симптомом возможна после n временных интервалов.

В качестве примера, на рис. 1 показано двумерное пространство, в котором симптомы трех разных причин (например, доступная память процессора, скорость входных данных, скорость выходных данных) образуют три кластера. Если в этом пространстве признаков имеется достаточное количество помеченных данных, можно изучить модель для классификации немаркированных точек в пространстве функций.

Одним из возможных результатов является попадание прогнозируемой точки на третьем временном шаге в кластер, представляющий аномальный симптом B. Если этот результат имеет большую вероятность, система должна предупредить, что аномалия с симптомом B будет происходить после трех временных шагов. Для того чтобы предсказать значения признаков, системе необходимо смоделировать статистические изменения шаблонов (паттернов) различных значений признаков. Сочетая классификацию симптомов аномалии и предсказание значения функции, система должна выполнять прогнозирование аномалий в сети, то есть выполнять классификацию симптомов аномалии для будущих данных

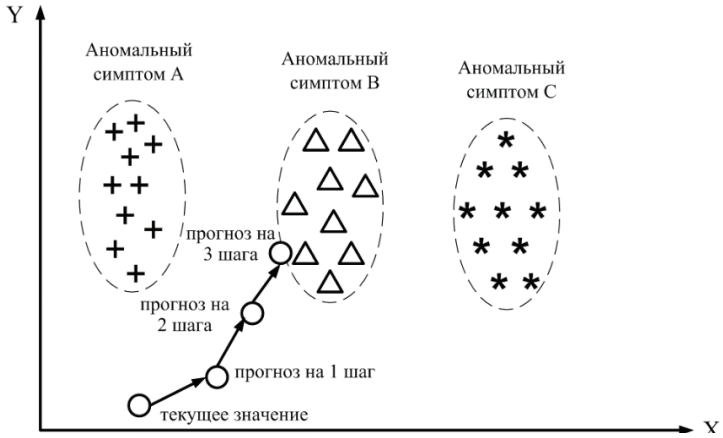

Рис. 1. Прогнозирование поведения КС

Для решения подобной задачи необходимо изучить различные аномальные симптомы из исторических данных (данные обучения), которые состоят из записей фиксированного набора атрибутов. Для системного мониторинга пространство функций X состоит из набора измерений уровня системы и уровня приложения. Это могут быть показатели уровня хоста, (такие, например, как доступная память, свободное время процессора и свободное место на диске и др.) или метрики на уровне компонентов (такие, например, как скорость поступления входных данных, скорость передачи выходных данных, время обработки данных, значения использования памяти компонента и т. д). Классификатор позволит определить, указывают ли данные полученные в результате прогноза на аномалию КС и получить оценки того, будет ли система иметь проблемную область в будущем [2, 16].

Понятие «Здоровье сети»

Будем характеризовать «здоровье КС» рядом системных показателей, характеризующих уровень обслуживания (Service Level Objectives, SLO), и соглашение об уровне предоставляемого сервиса SLA (*Service Level Agreement) рассматриваемой КС.* В качестве основных параметров, характеризующих SLA как правило используются:

- время отклика для конечного пользователя на запрос;
- время реакции сервера;
- задержка сигнала в сети.
- время ответа сервера на транзакцию;
- производительности диска,

и некоторые другие.

Будем характеризовать здоровье сети точкой (областью) в n-мерном пространстве основных параметров, характеризующих уровень обслуживания КС. В зависимости от того в какой области эта точка окажется можно диагностировать уровень «здоровья сети». Границы областей определяются на основе анализа нормально функционирующих сетей (что заранее известно) и сетей, содержащих проблемы, связанные с работоспособностью.

В итоге основной задачей определения "здоровья" сети является определения координат точки в n-мерном пространстве, характеризующем исследуемую КС для каждого из измерений.

Характеристики, используемые в процессе диагностики здоровья компьютерной сети

Чтобы упростить процесс диагностики здоровья компьютерной сети удобнее использовать один и тот же подход при определении координат. Например, можно для каждого из измерений выставлять баллы по результатам выполненного анализа (от 1 до 10), где 1 это худшее значение, а 10 это лучшее.

В качестве топологических характеристик для КС следует отметить продуманность реализованной топологии, которая может характеризоваться, например, связностью или же максимально длинной пути, нормированной к числу узлов сети. Топологической характеристикой можно также считать среднее число промежуточных звеньев, на пути следования пакетов и др. Выделив набор топологических характеристик, можно проставить оценки по каждой из них для исследуемой топологии.

Для изучения проблем, возникающих в КС и связанных с трафиком, требуется выполнить замер на одном из крупных каналов, например, в канале по которому проходит трафик из/в Интернета. Можно проанализировать трафик, записанный на крупных информационных ресурсах внутри сети, таких как базы данных, почтовые/веб-серверы и прочее.

При анализе "здоровья" сети требуется исследовать замеры трафика, которые позволят выявить наличие и характер аномальных участков, которые наблюдаются в трафике. Аномальный участок может присутствовать как в результате неверно сконфигурированной сети или несбалансированной нагрузки, так и в результате атак, которые были зафиксированы в исследуемом трафике. Оба эти фактора, обнаруженные в замерах трафика однозначно свидетельствуют о существовании проблем со "здоровьем" сети.

Для выявления проблем, связанных с функционированием оборудования, вычисления характеристик, описывающих проблемы с работой оборудования, требуется проанализировать участи, когда трафик в канале отсутствовал (частоту появления таких участков и их длительность). Такие участки свидетельствуют о перезагрузке промежуточных сетевых устройств или же обрывах каналов передачи данных.

Еще один индикатором проблем, связанных с функционированием оборудования является анализ числа переспросов на уровне TCP протокола, а также большое число ping-сообщений (без ответа), наблюдаемых в точке замера трафика.

Важным компонентом сбора информации о состоянии КС являются характеристики, основанные на анализе Log-файлов**.** При анализе log-файлов требуется определиться с типом Log-файлов, которые можно собирать параллельно с записью трафика, зафиксировать перечень ошибок, которые фиксируются в собираемом log-файле. Выполнив подобный анализ можно предоставить сводку по типам ошибок в системе, нормированную к общему числу сообщений, хранящихся в log-файле. Для построения качественных оценок для каждого из типов ошибок требуется выделить диапазоны, что позволит на основе бальной системы оценить интенсивность появления ошибок каждого типа.

Для диагностики здоровья КС требуется регулярно собирать характеристики, описанные выше. Для выбранного временного диапазона по результатам анализа собранных характеристик строится точка (координаты которой получены по результатам сделанных измерений), которая в дальнейшем может быть оценена с точки зрения принадлежности к участку нормального функционирования, либо участку, связанному с проблемами работы сети. Подробный анализ составляющих позволит найти совокупность характеристик, из-за которых "здоровье" сети было признано неудовлетворительным.

Для выделения участков нормального функционирования требуется выполнить предварительную обработку данных. В качестве инструмента предварительной обработки может быть использована кластеризация, например, на основе алгоритма k-means. Для получения значимых результатов кластеризации требуется выполнить квантование измеренных данных, чтобы снизить потенциальное число обнаруженных в данных кластеров. Желательно использовать уже размеченные данные, чтобы в результате кластеризации было понятно, какие группы измерений относятся к состояниям плохого/хорошего состояния "здоровья" компьютерной сети.

Прогнозирование "здоровья" компьютерной сети

Для прогнозирования состояния КС требуется подвергнуть данные предварительной обработке. При этом необходимо рассмотреть несколько уровней квантования качества функционирования КС, количество которых выбирается на основе экспертной оценки. Целесообразно для каждого показателя рассматривать небольшое число уровней квантования (как правило, не более 5), обеспечивающих приемлемую точность прогнозирования.

После предварительного квантования измеренных данных каждый показатель, содержащий в себе информацию о "здоровье" компьютерной сети, используется для прогнозирования на следующий момент времени. Для прогнозирования предлагается использовать алгоритм, построенный на отслеживании статистических зависимостей. Подобный алгоритм должен позволять строить прогноз для заданного исходного ряда с привлечением любых других связных последовательностей, в том числе и прогнозов этого же временного ряда, построенного любым другим методом прогнозирования. Следует отметить, что предложенный подход использует статистику зависимостей в данных и не ограничивается на поиске линейных зависимостей. Поэтому применение данного алгоритма представляется наиболее универсальным.

Как только для каждого из показателей работы компьютерной сети получен прогноз, требуется расположить полученное значение в пространстве, используемом для выполнения диагностики "здоровья" сети (с поправкой на сниженное число уровней квантования использованных показателей). В зависимости от того, к какой группе будет отнесено спрогнозированное состояние, будет сделан вывод о спрогнозированном "здоровье" компьютерной сети.

Таким образом, основной целью создания системы диагностики здоровья КС является разработка алгоритмов обнаружения, идентификации и прогнозирования аномальных

состояний, возникающих в процессе функционирования, на основе мониторинга системных показателей, поведения пользователей и анализа системных журналов, а также прогнозирование появления аномальных состояний и выявление причин их возникновения.

Интегральные сигнатуры состояния КС

Первой задачей является получение «отпечатка» состояния системы, которое приводит к нарушению SLO и может быть использовано для диагностики, кластеризации и участвовать в процессе поиска. Будем называть такое представление системы интегральной сигнатурой. С этой целью будем полагать, что:

1) состояние системы можно оценить в любой момент времени, например, изучив журнал логирования [3, 19].

2) можно непрерывно получать показатели системы, характеризующие низкоуровневые операции системы, такие, например, как использование центрального процессора, длину очереди, время ожидания ввода/вывода и другие. Подобная информация может быть получена с помощью различных инструментов, например, HP OpenView.

3) информация о значениях атрибутов «сырых» низкоуровневых показателей является ключом к построению сигнатур, характеризующих различные причины нарушения SLO.

Процесс ***измерения важности атрибутов*** происходит следующим образом. Входящие данные представляют собой векторы M низкоуровневых показателей системы и само состояние системы Y (нарушает SLO или нет). Разделим интервалы время на равные интервалы (например, по 5 минут) и определяем вектор M для каждого интервала. Каждый элемент m_i вектора M состоит из среднего значения показателя на протяжении интервала, а Y содержит дискретное значение – нарушено SLO или нет.

Функциональная схема системы диагностики здоровья КС

На рисунке 2 представлена обобщённая функциональная схема предлагаемой системы диагностики здоровья КС, а также предсказания рисков ухудшения здоровья. Опишем поэтапно процесс ее функционирования.

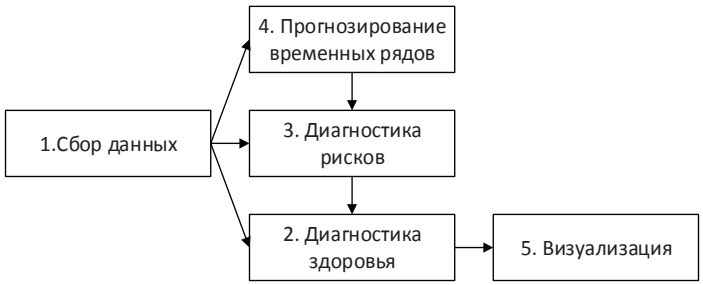

Рис. 2. Обобщённая функциональная схема системы диагностики здоровья КС

Модуль 1. Сбор данных. К данному модулю относятся программно-аппаратные средства, отвечающие за сбор и предварительную обработку данных. Важно отметить, что данный модуль реализует запись собираемых сведений в базу данных, а также предоставляет сведения по текущим измерениям Модулям 2, 3 и 4 для выполнения анализа поступающих сведений "на лету" (online).

Модуль 2. Диагностика здоровья. В модуле реализован алгоритм кластеризации собранных Модулем 1 данных, а также наносящий предсказанный вектор здоровья компьютерной сети (полученный от Модуля 3), на пространство уже подвергшееся кластеризации. В результате нанесения такого вектора Модуль 2 вычисляет принадлежность к одному или нескольким кластерам, присутствующим в предварительно размеченном пространстве. Предварительная разметка кластера выполняется на основе обработки «исторических данных» характеризующих работоспособность КУС «в прошлом».

***Модуль 3. Диагностика рис*ков.** Информация для анализа, выполняемого данным модулем, поступает как напрямую от Модуля 1, собирающего данные сенсоров компьютерной сети, так и от Модуля 4, формирующего прогнозы записанных временных рядов. Основная задача Модуля 3 состоит в формировании вектора прогноза здоровья сети на один или несколько временных интервалов в будущее и передачи этой информацию о вероятном будущем состоянии Модулю 2.

Модуль 4. Прогнозирование временных рядов. Модуль представляет собой набор программных библиотек для прогнозирования временных рядов. Информация для прогнозирования поступает от Модуля 1, а результаты сделанных прогнозов записываются в базу данных и являются вспомогательной информацией для получения прогноза "здоровья" компьютерной сети. Поэтому прогнозы временных рядов также доступны Модулю 3.

Модуль 5. Визуализация. Для построения траектории "здоровья" компьютерной сети и визуализации в двух/трёхмерном пространстве требуется использование методик снижения размерности рассматриваемого пространства. Данный алгоритмический подход реализован в виде программного обеспечения на уровне Модуля 5. Также здесь пользователю доступны численные данные, собираемые и прогнозируемые системой. При необходимости пользователь получает дополнительную информацию о диагнозе, т.е. к какому кластеру, из размеченных, текущее состояние здоровья сети наиболее близко, т.е. реализуется функционал предварительного анализа типа проблемы, с которой уже столкнулась сеть или может столкнуться в ближайшем будущем

Диагностика рисков нарушения «здоровья» компьютерной сети

Функциональная схема системы диагностики здоровья компьютерной сети представлена на рис. 3.

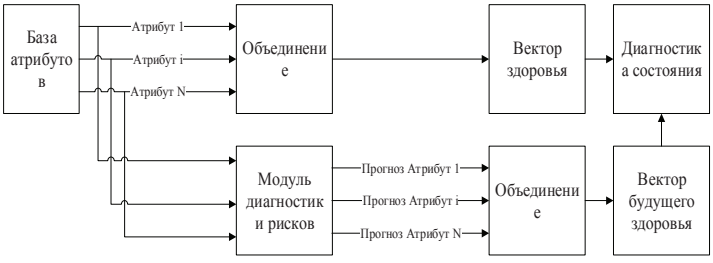

Рис. 3. Функциональная схема системы диагностики здоровья компьютерной сети

Как показано на рис. 3 основным элементом системы диагностики здоровья компьютерной сети является модуль диагностики рисков (МДР). Модуль диагностики рисков используется для формирования вектора «будущего здоровья" компьютерной сети. Для того чтобы получить значения вектора «будущего здоровья" компьютерной сети необходимо подать на вход МДР атрибуты, участвующие в формировании "здоровья" сети. По каждому из них формируется прогноз, после чего формируется Вектор «будущего здоровья и блок диагностики состояния оценивает, является ли полученный вектор к потенциально опасным состоянием компьютерной сети или нет.

МДР базируется на алгоритме прогнозирования вектора "здоровья" компьютерной сети в следующий (будущий) момент времени. Для диагностики будущего риска, его требуется оценить. Для этого и необходимо понять, в каком состоянии компьютерная сеть окажется в следующий либо другие моменты времени в будущем. Функциональная схема МДР показана на рис. 4.

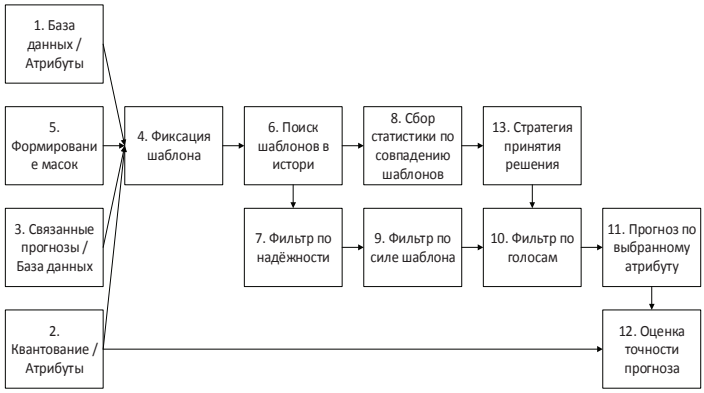

Рис. 4. Функциональная схема Модуля диагностики рисков

В качестве исходных данных, необходимых для оценки будущего состояния системы, выступают набор атрибутов, которые были получены от сенсоров компьютерной сети. Как показано выше параметрами алгоритма, который необходимо подобрать в процессе оптимизации являются:

– размер шаблона;
– размер связанного шаблона;
– размер истории;
– порог надёжности;
– порог силы сигнала;
– порог по голосованию.

Рассмотрим назначение и функции каждого блока МДР.

Исходными данными для работы алгоритма МДР являются сведения, хранимые в базе данных с атрибутами – блок 1. Для снижения вариативности данных и повышения точности результатов прогнозирования будущих состояний системы осуществляется предварительное квантование отобранных атрибутов в блоке 2. С этой целью для каждого из атрибутов следует выбирать своё индивидуальное правило и число уровней квантования, зависящие от специфики атрибута и его влиянии на оценку "здоровья" компьютерной сети. Для работы алгоритма предусмотрено использование для атрибутов прогнозов, создаваемых любыми другими методиками прогнозирования.

Для улучшения точности предусмотрена возможность использования прогноза искомого атрибута, полученного другими методами. Подобный подход позволяет повысить точность уже имеющихся прогнозов. В случае если уже имеющийся прогноз имеет низкую точность – его значение не будет учтено предлагаемым алгоритмом.

Фиксация шаблона, выполняемая в блоке 4, состоит в извлечении списка шаблонов, по которым будет производиться поиск схожих ситуаций уже имевших место в прошлом. Исходными данными являются атрибуты, прошедшие квантование, а также связанные прогнозы, подготовленные и сохранённые в отдельной базе данных. Список шаблонов извлекается по маскам, формируемым в Блоке 5. В качестве параметров здесь выступают «Размер шаблона», а также «Размер связанного шаблона». Связанный шаблон используется, если прогнозирование производится по нескольким временным рядам. Блок формирования масок 5 создаёт набор масок для извлечения шаблонов из квантованных атрибутов и/или связанных прогнозов. Список масок данный блок отсылает в Блок 4.

По результатам сформированного шаблона в Блоке 6 выполняется поиск совпадений *шаблона в «прошлой истории»*. В качестве исходных данных выступают квантованные атрибуты и/или связанные прогнозы. Для ограничения поиска задаётся параметр «Размер истории», на основании которого выбирается глубина поиска совпадающих шаблонов.

После того как поиск совпадений шаблонов завершён применяется фильтрация шаблонов с низкой надёжностью (7). В качестве исходных данных выступает список шаблонов, сформированный Блоком 4 с указанием числа совпадений по каждому из шаблонов в прошлом. Параметром, который используется на уровне данного блока является «Порог надёжности». Если число наблюдаемых в прошлом шаблонов не превышает данный порог, тогда шаблон не проходит фильтр и не рассматривается при принятии решения о будущем значении прогнозируемого атрибута.

В блоке 8 (Сбор статистики по совпадению шаблонов*)* осуществляется сбор и хранение статистики совпадения шаблонов в прошлом путем формирования таблицы. Элементы таблицы указывают, какое число совпадений было зафиксировано по каждому шаблону, а также определяется временной сдвиг для каждого совпавшего шаблона относительно момента прогноза. Все собранные сведения поступают на вход блока «Стратегия принятия решения» о прогнозе.

Блок 9 (Фильтр по «силе» шаблона) реализует фильтрацию по силе шаблона, который прошёл порог, установленный Блоком 7. Под силой шаблона понимается степень выраженности одного из исходов, которые наблюдаются по рассматриваемому шаблону. Это, например, может быть величина, характеризующая, насколько чаще встречается один из исходов по отношению к следующему в списке по частоте встречаемости. Данный параметр задаётся в относительных единицах и называется «Порог силы сигнала». Шаблоны, прошедшие фильтр Блоков 8 и 9, подаются на окончательный фильтр 10, определяющий с каким перевесом по голосам шаблоны должны участвовать в принятии решения. Данный перевес является настраиваемым параметром и задаётся в виде относительной величины.

По мере поступления новых сведений о значениях атрибутов в блоке 12 формируется оценка точности сформиро-

ванных Блоком 11 Прогнозов по выбранному атрибуту. Результаты оценки точности могут быть использованы для переобучения (повторного выбора параметров алгоритма), если качество ранее сделанных прогнозов не отвечает требованиям, предъявляемым к системе.

Стратегия принятия решения (блок 13) определяет правило, по которому происходит голосование в блоке 10. Информацию о том, какое правило требуется использовать для выбранного атрибута, определяют на этапе оптимизации работы алгоритма по заданной целевой функции. Блок 13 «Стратегия принятия решения» информирует о принятом в качестве прогноза решении по атрибуту. При необходимости данное решение отображается и записывается в вектор будущего состояния "здоровья" компьютерной сети. Затем этот вектор используется «Модулем диагностики здоровья сети», а результаты такой диагностики будущего "здоровья" компьютерной сети отображаются «Модулем визуализации».

Выводы

Предложенное алгоритмическое и программное решение может быть использовано в системах мониторинга качества функционирования компьютерных систем, диагностики возникающих проблем, а также раннего обнаружения (на основе алгоритмов предсказания) появления рисков снижения качества функционирования компьютерной сети.

Полученные результаты предназначены для использования в качестве алгоритмической базы для создания программно-аппаратных продуктов диагностики функционирования и предсказания аномальных состояний и неисправностей компьютерных систем.

Литература

1. *M. Shatnawi and M. Hefeeda*. Real-time failure prediction in online services // IEEE Conference on Computer Communications (INFOCOM), 2015. С. 1391-1399.
2. Рекомендация МСЭ-Т М.3342. Указания по определению шаблонов представления SLA, 2006.
3. *Шелухин О.И., Рябинин В.С., Фармаковский М.А*. Обнаружение аномальных состояний компьютерных систем средствами интеллектуального анализа данных системных журналов. Вопросы кибербезопасности №2(26), 2018. DOI: 10.21681/2311-3456-2018-2-33-43
4. *F. Salfner, M. Lenk, and M. Malek*. A survey of online failure prediction methods // ACM Computing Surveys (CSUR), vol. 42, no. 3, 2010. С. 10.
5. *F. Salfner, M. Schieschke, and M. Malek*. Predicting failures of computer systems: A case study for a telecommunication system // Proceedings 20th IEEE International Parallel & Distributed Processing Symposium, 2006.
6. *V. Balaji and V. Duraisami*. Cluster based packet loss prediction using tcp ack packets in wireless network // (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010. С. 2313-2315.
7. *Ira Cohen, Steve Zhang, Moises Goldszmidt, Julie Symons, Terence Kelly*. Capturing, Indexing, Clustering, and Retrieving System History. SOSP'05, 2005, Brighton, United Kingdom. Copyright 2005 ACM 1-59593-079-5/05/0010.
8. *S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox*. Ensembles of models for automated diagnosis of system performance problems. DSN, 2005.
9. *Z. Yang, M. Kitsuregawa*. LPI-SPAM: An Improved Algorithm for Mining Sequential Pattern // Proc. of Int'l Special Workshop on Databases for Next Generation Researchers in conjunction with ICDE'05, 2005. С. 8-11.
10. *M.J. Zaki.* SPADE: An Efficient Algorithm for Mining Frequent Sequences // Machine Learning Journal, Vol. 42(1/2), 2001, с. 31-60.
11. Mohammed J. Zaki. SPADE: An Efficient Algorithm for Mining Frequent Sequences [Текст]; Machine Learning, №42, 2001. С. 31-60.
12. *Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu*. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach // IEEE Transactions On knowledge and data engineering, vol. 16, no. 10, 2004.
13. *R.Agrawal and R.Srikant*. Mining sequential patterns // Proceedings of the Eleventh International Conference on Data Engineering, 1995.
14. *Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen.* A General Model for Sequential Pattern Mining with a Progressive Database Publication // IEEE Transactions On Knowledge And Data Engineering, Vol. 20. No. 9, 2008.
15. *S. Abbasghorbani and R. Tavoli.* Survey on sequential pattern mining algorithms // 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2015, pp. 1153-1164. doi: 10.1109/KBEI.2015.7436211
16. *Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage-Uday Kiran, Yun-Sing Koh, and Rincy Thomas*. 2017. A survey of sequential pattern mining. Data Science and Pattern Recognition 1, 1 (2017), pp. 54-77.
17. *Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu*. A Survey of Parallel Sequential Pattern Mining. ACM Trans. Knowl. Discov. Data. 0, 1, Article 00 (August 2018), 33 pages. https://doi.org/0000001.
18. *Xiaohui Gu*. Online Anomaly Prediction for Robust Cluster Systems // IEEE 25th International Conference on Data Engineering. March 2009, pp. 1000-1011. DOI: 10.1109/ICDE.2009.128
19. *Шелухин О.И., Рябинин В.С.* Обнаружение аномалий больших данных неструктурированных системных журналов // Вопросы кибербезопасности. 2019. №2(30). С. 36-41. DOI 10.21681/2311-3456-2019-2-36-41

MONITORING AND DIAGNOSTICS OF ANOMALOUS STATES IN A COMPUTER NETWORK BASED ON THE STUDY OF "HISTORICAL DATA"

Oleg I. Sheluhin, *MTUCI, Moscow, Russia,* ***sheluhin@mail.ru***
Andrey V. Osin, *MTUCI, Moscow, Russia,* ***osin_a_v@mail.ru***
Denis V. Kostin, *MTUCI, Moscow, Russia,* ***d.v.kostin@mail.ru***

Abstract
This paper proposed to characterize the "health of a computer network" by a set of system metrics that characterize the Service Level Objectives and Service Level Agreement of the computer network. The necessary parameters (attributes, signatures) that determine the state of a computer network can be extracted from historical data and used to automatically cluster and search for similar problems in the past based on similarities. The database of historical events allows to find and compare the current behavior of the system with similar previously encountered problems that were observed in the past. To solve this problem, it is necessary to study various abnormal symptoms from historical data at the training stage. To predict "future" symptoms, it is necessary to model statistical changes in patterns of different attribute values. The functional diagram of health diagnosis and risk prediction has been proposed. The paper studies the characteristics for determining the health of a computer network. Combining the classification of anomaly symptoms and prediction, the diagnostic system must predict network anomalies based on the classification of anomaly symptoms for future data. An algorithmic and software solution can be used to monitor the quality of computer systems, for early detection (based on prediction algorithms) and to identify various problems that reduce the quality of a computer network.

Keywords: *anomaly states, computer network, forecasting, machine learning, data mining, monitoring system metrics, clustering, sequential analysis; pattern.*

References

1. M. Shatnawi and M. Hefeeda. (2015). Real-time failure prediction in online services. *IEEE Conference on Computer Communications (INFOCOM)*, pp. 1391-1399.
2. ITU-T Recommendation M.3342. (2006). *Guidelines for Defining SLA Presentation Templates.*
3. Sheluhin O.I., Ryabinin V.S., Farmakovsky M.A. (2018). Detection of abnormal conditions of computer systems by means of data mining system logs. *Cybersecurity Issues.* No. 2 (26). DOI: 10.21681 / 2311-3456-2018-2-33-43
4. F. Salfner, M. Lenk, and M. Malek. (2010). A survey of online failure prediction methods. *ACM Computing Surveys (CSUR)*, vol. 42, no. 3, pp. 10.
5. F. Salfner, M. Schieschke, and M. Malek. (2006). Predicting failures of computer systems: A case study for a telecommunication system. *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium.*
6. V. Balaji and V. Duraisami. (2010). Cluster based packet loss prediction using tcp ack packets in wireless network. *(IJCSE) International Journal on Computer Science and Engineering.* Vol. 02, No. 07, pp. 2313-2315.
7. Ira Cohen, Steve Zhang, Moises Goldszmidt, Julie Symons, Terence Kelly. (2005). Capturing, Indexing, Clustering, and Retrieving System History. *SOSP'05, Brighton, United Kingdom.* Copyright 2005 ACM 1 59593 079 5/05/0010.
8. S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox. (2005). Ensembles of models for automated diagnosis of system performance problems; DSN.
9. Z. Yang, M. Kitsuregawa. (2005). LPI-SPAM: An Improved Algorithm for Mining Sequential Pattern. *Proc. of Int'l Special Workshop on Databases for Next Generation Researchers in conjunction with ICDE'05*, pp. 8-11.
10. M.J. Zaki. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning Journal,* Vol. 42(1/2), pp. 31-60.
11. Mohammed J. Zaki. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, no. 42, pp. 31-60.
12. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu. (2004). Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. *IEEE Transactions On knowledge and data engineering*, vol. 16, no. 10.
13. R.Agrawal and R.Srikant. (1995). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering.*
14. Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen. (2008). A General Model for Sequential Pattern Mining with a Progressive Database Publication. *IEEE Transactions On Knowledge And Data Engineering*, Vol. 20, No. 9.
15. S. Abbasghorbani and R. Tavoli. (2015). "Survey on sequential pattern mining algorithms". *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI),* Tehran, pp. 1153-1164. doi: 10.1109/KBEI.2015.7436211
16. Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage-Uday Kiran, Yun-Sing Koh, and Rincy Thomas. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition* 1, 1, pp. 54-77.
17. Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu. (2018). A Survey of Parallel Sequential Pattern Mining.;ACM Trans. Knowl. Discov. Data. 0, 1, Article 00 (August 2018), 33 pages. https://doi.org/0000001
18. Xiaohui Gu. (2009). Online Anomaly Prediction for Robust Cluster Systems. *IEEE 25th International Conference on Data Engineering*. March 2009. Pages 1000-1011. DOI: 10.1109/ICDE.2009.128.
19. Sheluhin O.I., Ryabinin V.S. (2019). Detection of large data anomalies in unstructured syslogs. *Cybersecurity issues*. No. 2 (30), pp. 36-41. DOI 10.21681 / 2311-3456-2019-2-36-41

Information about authors:
Oleg I. Sheluhin, *doctor of technical sciences, professor, head of the Department of Information Security, MTUCI, Moscow, Russia*
Andey V. Osin, *PhD, MTUCI, Moscow, Russia*
Denis V. Kostin, *graduate student, MTUCI, department of information security, Moscow, Russia*